VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ KATEDRA

Didelių duomenų srautų analizė, anomalijų aptikimas

Big data analysis, detection of anomalies

Kursinis darbas

Atliko: 3 kurso 3 grupės studentas
Jokūbas Rusakevičius (parašas)

Darbo vadovas: dr. Vytautas Valaitis (parašas)

Vilnius – 2018

TURINYS

ĮVADAS 3
Problematika 3
Darbo tikslas ir uždaviniai 4

1. DIDIEJI DUOMENYS 5
1.1. Didžiųjų duomenų didieji „V“ 6
1.1.1. Keikis - „Volume“ 6
1.1.2. Greitis - „Velocity“ 6
1.1.3. Įvairovė - „Variety“ 6
1.1.4. Teisingumas - „Veracity“ 7
1.1.5. Vertė - „Value“ 7
1.2. Pritaikymas 7

2. DIDŽIŲJŲ DUOMENŲ ANALIZAVIMAS 8

3. „MACROBASE“ ANALITINIS ĮRANKIS 9

4. DUOMENŲ ANALIZĖS EKSPERIMENTAS NAUDOJANT „MACROBASE“ ANALITINĮ ĮRANKĮ 10
4.1. Duomenų rinkinys 10
4.1.1. Eksperimentui pritaikyti duomenys 10
4.1.1.1. Duomenų aprašymas 10
4.1.1.2. Eksperimentui atrinkti naudingi SMART atributai 10
4.1.1.3. Duomenų paruošimas 11
4.2. Eksperimentui naudotos aplinkos aprašimas 12
4.2.1. Virtuali mašina eksperimentui 12
4.2.1.1. Virtualios mašinos paruošimas 12
4.2.2. „MacroBase GUI“ analitinio įrankio su grafine naudotojo sąsaja paruošimas 13
4.2.2.1. „MacroBase GUI“ reikalingi papildomi paketai 13
4.3. Eksperimento vykdymo eiga 13
4.3.1. Eksperimento eiga su „Backblaze“ duomenimis 14
4.3.2. Eksperimento eiga su jungtiniais duomenimis 14
4.3.2.1. Eksperimentas: mėnesio dienos, kai diskų temperatūros yra neįprastai aukštos 14
4.3.2.2. Eksperimentas: darbas su dideliais CSV failais 14
4.4. Eksperimento rezultatų analizė 15
4.4.1. Eksperimentų su „Backblaze“ duomenimis gauti rezultatai 15
4.4.2. Eksperimentų su jungtiniais duomenimis gauti rezultatai 16
4.4.2.1. Dienų, kai diskų temperatūros yra išskirtinai aukštos eksperimento rezultatai 16
4.4.2.2. Darbo su dideliais CSV failais eksperimento rezultatai 16

REZULTATAI IR IŠVADOS 17

LITERATŪRA 18

SANTRUMPOS 21

PRIEDAI 21
1 priedas. „Backblaze“ duomenų SMART. atributų reikšmės 22
2 priedas. „Backblaze“ CSV failų apjungimo BASH skriptas 25
3 priedas. „MacroBase GUI“ pagrindinis langas 26
4 priedas. Diskų gedimų numatymo eksperimento rezultatai 27

5 priedas. Diskų gedimų numatymo eksperimento rezultatų statistikos 33

Įvadas

Šis darbas yra Programų Sistemų studijų trečio kurso privalomas kursinis darbas apie anomalijų aptikimą didžiuosiuose duomenyse (angl. *Big Data*) bei anomalijų aptikimą naudojant atviro kodo analitinį įrankį „MacroBase“.

Problematika

Remiantis 2001 metais „Gartner“ pateiktu apibrėžimu (iki šiol laikomu pagrindiniu (angl. *go-to*)), didieji duomenys - tai duomenys, kurie turi didelę įvairovę (angl. *variety*), yra renkami nuolat didėjančiais kiekiais (angl. *volumes*) ir generuojami vis didėjančiais greičiais (angl. *velocity*), šis apibrėžimas dar vadinamas tryjų „V“ [Lan01]. Paprastai, didžiuosius duomenis galima apibrėžti kaip duomenų rinkinius tokius didelius, kad tradiciniai programiniai duomenų apdorojimo įrankiai nesugeba jų sugauti, tvarkyti, organizuoti, apdoroti ar su jais dirbti priimtiname laiko intervale [SMR12], jie yra paprasčiausiai per dideli ir per daug sudėtingi. Tačiau šie duomenys turi milžinišką potencialą ir gali būti panaudojami sprendžiant verslo ir kitas problemos, kurių sprendimas iki šiol buvo neįmanomas.

Generuojant ir saugant milžiniškus kiekius duomenų, natūraliai, užfiksuojami tokie duomenys, kurie išsiskiria ir yra nebūdingi duomenų rinkiniui. Duomenų vienetai kurie yra nukrypę nuo kitų duomenų rinkinyje yra vadinami anomalijomis. Anomalijų aptikimas yra procesas, kurio metu yra aptinkama ir identifikuojama anomalija arba išskirtis (angl. *outlier*) duomenų rinkinyje [Tec18a]. Anomalijos yra retas reiškinys, tačiau jų egzistavimas gali reikšti didelį pavojų taikomajai sistemai ar jos naudotojams. Šio darbo metu bus tiriamas anomalijų aptikimas remiantis iš „Backblaze“ 2018 metų (pirmą ketvirtį) duomenų centruose esančių diskų surinkta informacija [Bac18].

Dėl įvairių priežasčių renkamų, saugomų ir operuojamų duomenų kiekiai nuolatos didėja ir netgi gerokai lenkia žmogaus sugebėjimą juos apdoroti ar analizuoti. Didžiosios socialinių tinklų kompanijos Twitter, Facebook ir LinkedIn 2015-2016 metais pranešė kiekviena atskirai fiksuojanti iki 12 milijonų įvykių per sekundę [Ast16; PFT+15; Woo15]. Taip pat negalima pamiršti vis labiau plintančių ir didelius kiekius duomenų generuojančių automatizuotų duomenų šaltinių („Dalykų Interneto“ (angl. *Internet of Things* arba *IoT*)). Be to, tokie palankūs veiksniai kaip kylantis automatizuotų duomenų šaltinių populiarumas, pinganti techninė įranga, išvystyti komunikaciniai tinklai bei mažėjančios duomenų saugojimo kainos paskatino dešimčių milijardų dolerių komercines investicijas šių technologijų vystimui [MCB+15]. Dėl šių ir kitų veiksnių numatoma, kad kiekvienais metais bendras duomenų kiekis išaugs po 40% [EMC14], o iki 2020 metų pranašaujama, kad bendras pasaulinis duomenų kiekis peržengs 40 zetabaitų (4×10^{22} baitų) ribą [HA17].

Didžiųjų duomenų laikomos informacijos paslėpta nauda ir svarba yra visuotinai pripažįstama, tačiau ši informacija nėra lengvai išgaunama. Duomenų peržiūra ir analizė sudaro labai didelį krūvį tiek analitikui, tiek analitiniams įrankiams. Fizinė duomenų peržiūra yra paprasčiausiai neįmanoma, o nuolatos didėjantys duomenų kiekiai vis labiau atskiria dėmesio reikalaujančius

duomenis ir ribotą dėmesį turintį analitiką. Net ir aukščiausios kvalifikacijos analitikai praneša panaudojantys vos iki 6% jų surenkamų duomenų [BGR+17]. Todėl atsiranda iššūkis prioritizuoti žmogaus dėmesį. Nors žmogui yra neįmanoma peržiūrėti visų šių duomenų, tačiau kompiuteriai ir/ar mašinos gali. Informacinės sistemos labiau nei bet kada turi filtruoti, akcentuoti, jungti, grupuoti pateiktus duomenis, jiems suteikti kontekstą ir rodyti naudotojui tik ribotą, svarbią bei apibendrintą informaciją. Visa rodoma, bet nereikalinga informacija reikalauja ir eikvoja žmogaus dėmesį [Sim71].

Standfordo universitetas kartu su Masačusetso technologijos institutu 2017 metais paskelbė kuriantys naują atviro kodo, ne tik didžiųjų, bet ir greitųjų duomenų (angl. *Fast Data*) analitinį paieškos įrankį. „MacroBase“ pagrindinis uždavinys yra žmogaus dėmesio prioritizavimas. Vienas iš „MacroBase“ šio uždavinį sprendimų yra sugeneruoti didžiausio dėmesio reikalaujančią supaprastintą išvestį, kurios neįprastus duomenų vienetus „MacroBase“ padeda aiškinti pagal duomenų atributus [BGM+17; BGR+17]. „MacroBase“ yra naujas ir modernus analitinis įrankis, pateikiantis inovatoriškų sprendimų vis didėjančių ir greitėjančių duomenų srautų analizei atlikti bei galintis grąžinti tikslius rezultatus dirbdamas 2 milijonų įvykių per sekundę greičiu per užklausą per branduolį, dėl to, šiame darbe bus plačiau nagrinėjamas bei eksperimentai atliekami naudojant būtent šį įrankį.

Darbo tikslas ir uždaviniai

Šio darbo **tikslas** - palyginti iki šiol naudotas didelių duomenų analizavimo technologijas ir pasinaudoti „MacroBase“ duomenų analizavimo ir anomalijų aptikimo įrankį anomalijų aptikimui.

Darbui iškelti **uždavyniai**:

1. Paaiškinti kas yra „Didieji Duomenis“.
2. Surasti ir palyginti dabar naudojamus anomalijų dideliuose duomenyse aptikimo įrankius.
3. Išanalizuoti „MacroBase“.
4. Paruošti eksperimentui reikalingą įrašų rinkinį iš „Backblaze“ kiekvieną ketvirtį skelbiamų duomenų.
5. Įsidiegti ir paruošti darbui „MacroBase“ analitinį įrankį.
6. Atlikti eksperimentus ir aptikti anomalijas paruoštuose duomenyse.
7. Pateikti galutines eksperimento išvadas.

1. Didieji duomenys

Pasak MIT Media Lab direktoriaus Joi Ito, duomenų užrašymas ant popieriaus lapo dar visai neseniai buvo vienintelis būdas rinkti informaciją. Žmogus sugalvotas idėjas ir mintis užrašydamas ant popieriaus lapo jas paversdavo žiniomis. Tačiau dabartinė didžiųjų duomenų situacija yra kitokia. Priešingai nei seniau, surenkamų duomenų kiekiai yra milžiniški, tačiau jie nėra žinios, tol kol jų nepradedama nagrinėti ir analizuoti. Tik pradėjus analizuoti duomenis atliekant su jais įvairias transformacijas, galima pastebėti, kad gaunama įdomi ir netgi svarbi informacija[Smo14].

Terminas „Didieji duomenys“ vartojamas apibūdinant procesą naudojamą, kai tradicinės duomenų rinkimo (angl. *data mining*) ir tvarkymo metodikos nebegali įžvelgti ar atskleisti duomenyse esančios prasmės [Tec18b]. Nors pirmieji dideli duomenų įrašų rinkiniai yra datuojami jau 1960-1970 metais [Ora18b], pats terminas buvo pradėtas naudoti tik nuo 1990 metų, o jo autoriumi yra laikomas ar bent už termino išgarsinimą yra dėkojama John Mashley [Loh13; Mas98; Pre13]. Tačiau kaip didžiųjų duomenų „dydis“ nuolatos plečiasi ir kinta, taip ir jų apibrėžimas, bei apibrėžimui naudojamų „V“ kiekis (1.1 poskyris).

Didieji duomenys apima visų trijų struktūrizacijos lygių duomenis: struktūrizuotus, iš dalies struktūrizuotus ir nestruktūrizuotus, tačiau dižiausias didžiųjų duomenų dėmesys yra skiriamas nestruktūrizuotiems duomenims [DS16]. Šie skirtingo struktūrizacijos lygio duomenys pasižymi tokiomis savybėmis:

- **Struktūrizuoti duomenys** - tai duomenys, laikomi jiems skirtuose fiksuotuose laukuose, failuose ar įrašuose. Struktūrizuoti duomenys priklauso nuo jiems priskirto duomenų modelio, kuris nurodo kokie, kokio tipo (skaitiniai duomenys, valiutos, alfabetiniai duomenys, vardai, datos, adresai) ir kokių apribojimų (simbolių skaičius, terminų apribojimas, pvz.: Ponas, Ponia, Prof., Doc. ir kt) laukai bus saugomi. Šio tipo duomenys yra lengvai operuojami, pridedami ir analizuojami. Dėl didelių duomenų saugojimo kaštų tai ilga laiką buvo vienintelis sprendimas duomenų saugojimui - viskas, ko negalima optimaliais struktūrizuoti, laikoma popieriniame formate [Web18].

- **Iš dalies struktūrizuoti duomenys** [Bun97] - tai tarpinis variantas tarp struktūrizuotų ir nestruktūrizuotų duomenų. Duomenys neturi jiems priskirtos formalios duomenų modelio struktūros susietos su duomenų bazėmis ar kitomis duomenų lentelių formomis, tačiau turi žymes ar kitas žymėjimo priemones įrašų atskyrimui ir įrašų hierarchijos sukūrimui. Todėl iš dalies struktūrizuoti duomenys yra dar žinomi kaip save apibūdinančios (angl. *self-describing*) struktūros. Iš dalies struktūrizuotuose duomenyse įrašai gali priklausyti vienai klasei ir būti grupuojami kartu, tačiau turėti skirtingus atributus, o pati atributų išdėstymo tvarka nėra svarbi. XML ir kitos žymių kalbos yra iš dalies struktūrizuotų duomenų pavyzdžiai [Web18].

- **Nestruktūrizuoti duomenys** - tai duomenys, kuriems skiriamas didžiausias didžiųjų duomenų dėmesys. Šio tipo duomenys neturi jiems priskirto duomenų modelio ar taisyklių pagal, kurias jie būtų organizuojami, ir jų neišeina klasifikuoti pagal žmogui skaitomus

požymius. Tai daugiaprasmiai duomenys, kurių sudėti tvarkingai į vieną „dėžutę“ yra neįmanoma: nuotraukos, paveikslėliai, vaizdo įrašai, transliuojami instrumentiniai duomenys, PDF dokumentai, pristatymų skaidrės, elektroniniai laiškai, tinklaraščiai, teksto dokumentai, knygos, medicinos įrašai ir kt. Nors kiekvienas prieš tai išvardytas duomenų tipas atskirai yra struktūrizuotas, tačiau jų rinkinys vis tiek laikomas nestruktūrizuotais duomenimis [Web18]. Yra apskaičiuota, kad apie 80% visų duomenų yra nestruktūrizuoti [Sch16].

Toliau šiame skyriuje bus aprašomi didžiuosius duomenis apibūdinantys penki „V“ (1.1 poskyris. Galiausiai bus rašoma apie didžiųjų duomenų pritaikymą (1.2 poskyris).

1.1. Didžiųjų duomenų didieji „V“

Originaliai 2001 metais pateiktas didžiųjų duomenų apibrėžimas susidėjo iš trijų „V“: įvairovės (angl. *variety*), kiekio (angl. *volume*) ir greičio (angl. *velocity*) [Lan01]. Tačiau per paskutinius kelis metus šis apibrėžimas buvo papildytas dviem naujais „V“: duomenų teisingumu (angl. *veracity*) ir verte (angl. *value*) [Mar14]. 2016 metais pateiktas atnaujintas didžiųjų duomenų apibrėžimas: didžiuosius duomenis sudaro informacijos rinkiniai charakterizuojami tokių aukštų duomenų kiekių, greičių ir įvairovės, kad yra reikalingos specifinės technologijos ir analitiniai metodai vertės iš tos informacijos išgavimui [MGG16]. Toliau šiame poskyryje bus aprašomas kiekvienas „V“ (Kiekis (1.1.1 punktas), greitis (1.1.2 punktas), įvarovė (1.1.3 punktas), teisingumas (1.1.4 punktas) ir vertė (1.1.5 punktas)) atskirai.

1.1.1. Keikis - „Volume“

Kiekis nusako didžiulį kiekį duomenų, kurie yra sugeneruojami kiekvieną sekundę. „FaceBook“ kiekvieną dieną sugeneruoja 4 petabaitus ($1PB = 10^{15}B$) naujų duomenų, vien mygtukas „Patinka“ yra paspaudžiamas virš 4 milijonų kartų per minutę, o naujų kiekvieną dieną įkeliamų nuotraukų skaičius siekia net 350 milijonų [Smi18]. Sugeneruojamų duomenų kiekis per minutę beveik prilygsta visiems iki 2008 metų surinktiems duomenims. Tačiau pasinaudojus didžiųjų duomenų technologija ir išskirstytą (angl. *distributed*) sistemą, kur duomenų dalys sujungtos programinės įrangos yra laikomos skirtingose vietose, yra įmanoma juos saugoti ir jais naudotis.

1.1.2. Greitis - „Velocity“

Gretis nusako greitį, kuriuo yra generuojami nauji duomenys, ir greitį kuriuo duomenys yra perduodami. „CERN“ „LHC“ dalelių greitintuvas generuoja 1 megabaitą neapdorotų duomenų per įvykį, o per vieną sekundę yra užregistruojami net 600 milijonų įvykių (iš viso 600 TB/s) [Cer18]. Didžiųjų duomenų technologijos sudaro sąlygas analizuoti duomenis, kol jie yra generuojami taip niekada nepatalpinant jų į duomenų bazę.

1.1.3. Įvairovė - „Variety“

Įvairovė nusako skirtingus duomenų tipus. Dauguma duomenų dabar, ne taip kaip praeityje, yra nestruktūrizuoti, kas padaro juos sunkiai patalpinamus į duomenų lenteles. Pasinaudojus

didžiųjų duomenų technologija galima sujungti skirtingų rūšių duomenis kartu su tradiciniais struktūrizuotais duomenimis.

1.1.4. Teisingumas - „Veracity“

Teisingumas nusako duomenų netvarkingumą ir patikimumą. Dėl skirtingų didžiųjų duomenų formų, kokybė ir tikslumas yra beveik nevaldomi. Tačiau didieji duomenys ir analitinės technologijos sugeba dirbti su šio tipo duomenimis, dažnai kompensuodamos duomenų netikslumą ir nekokybiškumą dideliais duomenų kiekis.

1.1.5. Vertė - „Value“

Vertė nusako iš didžiųjų duomenų gaunamą naudą ir dažnai yra laikomas svarbiausiu iš visų „V“. Yra svarbu dirbant su didžiaisiais duomenimis paversti juos į tam tikra vertę. Tai nebuvo vienas iš originaliai apibrėžtų „V“, tačiau buvo įvardintas dėl didelių darbo su didžiaisiais duomenimis kaštų ir nevisada aiškios gaunamos naudos.

1.2. Pritaikymas

Didžiuosius duomenis pritaikymas yra galimas beveik bet kokioje srityje, kur yra generuojami dideli kiekiai duomenų. Didžiųjų duomenų industrija tokia didelė, kad 2010 metais ji buvo verta apie 100 milijardų dolerių ir augo po beveik 10% per metus [Eco10]. Didieji duomenys yra populiarūs, nes iš jų gaunamos žinos yra plačiai pritaikomos. Todėl didžiųjų duomenų pritaikymo pavyzdžių šiais laikais galima rasti visur:

- Gamintojai gauna vieną didžiausių naudų. Naudojant didžiuosius duomenis gamintojai gali tobulindami savo tiekimo planus, gaminti pagal numatytą paklausą ir pasiekti beveik nulinį tiekimo laiką.
- Medicina didieji duomenys yra naudojami siekiant pritaikyti suasmenintą mediciną [HC16].
- Edukacinės institucijos kaip universitetai pradėjo kurti su duomenų analize susijusias studijų programas 2011 metais paskelbtai 1,5 milijonų aukštos kvalifikacijos duomenų analitikų trūkumo [MCB+11] paklausai patenkinti.
- Žiniasklaida naudoja naudoja didžiųjų duomenų analizę pateikdamos suasmenintas, kryptingas reklamas.
- Kt.

Su vis didėjančiais duomenų kiekiais ir generavimo greičiais didėja ir duomenyse esančių žinių potencialas. Remiantis 2014 metų duomenimis, vienas trečdalis visų duomenų yra saugomi skaitiniu ir/ar raidiniu (angl. *alphanumeric*) tekstų arba nuotraukų pavidalu [Hil14], o tai formatas naudingiausias duomenų pritaikymui. Taip pat, didelis potencialas slypi didžiųjų duomenų analizėm nenaudojamuose vaizdo ir audio bei kituose duomenyse.

2. Didžiųjų duomenų analizavimas

3. „MacroBase“ analitinis įrankis

4. Duomenų analizės eksperimentas naudojant „MacroBase“ analitinį įrankį

Eksperimentui atlikti buvo pasirinktas „MacroBase GUI“ analitinio įrankio „MacroBase“ grafinė naudotojo sąsaja. CSV tipo failai buvo pasirinktas duomenų šaltinio tipas atliekant eksperimentą.

4.1. Duomenų rinkinys

Šiame poskyryje aprašyti eksperimentui pritaikyti „Backblaze“ standžiųjų diskų stebėjimo duomenys [Bac18].

4.1.1. Eksperimentui pritaikyti duomenys

Eksperimentui bus naudojami „Backblaze“ duomenų centruose esančių standžiųjų diskų atliekamo darbo stebėjimų duomenys. „Backblaze“ nuo 2013 metų kiekvieną ketvirtį paviešina surinktus duomenis stebint jų duomenų centruose esančius kietuosius diskus. Kiekvieno ketvirčio duomenys yra pateikiami CSV failais - vienas CSV failas vienai dienai. Vidutinis vienos dienos įrašų eilučių skaičius faile - 100 tūkst; dydis - 28MB. Eksperimentui buvo pasirinktas, paskutinis paviešintas, 2018 metų pirmo ketvirčio duomenų rinkinys [Bac18] - 90 CSV failų.

4.1.1.1. Duomenų aprašymas

Kiekvienas „Backblaze“ duomenų centrų dienos įrašų CSV failas yra sudarytas iš duomenų, kurių didžiąją dalį sudaro „S.M.A.R.T.“ [Wik18] atributai:

- ***Date*** - įrašo įrašymo data, užrašoma yyyy-mm-dd formatu (dėl grupavimo į failus pagal dienas, visi įrašai viename faile turės tą pačią datą).
- ***Serial Number*** - gamintojo priskirtas kietojo disko serijos numeris.
- ***Model*** - gamintojo priskirtas kietojo disko modelio numeris.
- ***Capacity Bytes*** - disko dydis baitais.
- ***Failure*** - įvykusios klaidos žymėjimui skirtas atributas. „0“, jei kietasis diskas dirba korektiškai; „1“, jei tai buvo paskutinė diena, kai diskas buvo naudojamas prieš sugesdamas.
- ***SMART*** atributai - 100 stulpelių duomenų iš kurių 50 yra neapdorotų ir 50 normalizuotų duomenų stulpelių (atributų reikšmės yra aprašytos prieduose (Priedas nr. 1)).

4.1.1.2. Eksperimentui atrinkti naudingi SMART atributai

Didelė dalis „Backblaze“ duomenų rinkinio įrašų atributų nėra naudingi atliekamam eksperimentui. Todėl buvo pasirinkti keli naudingi atributai atmetant, tokius atributus kurie yra specifiniai gamintojams arba nėra suprantami, be disko konteksto. Buvo palikti trivialūs atributai

nurodantys disko temperatūrą ir bendrą valandinį disko darbo laiką, bei 5 atributai, kuriuos „Backblaze“ naudoja numatyti diskų sugedimus [Kle16]: perskirstytų sektorių skaičių, neištaisomų klaidų skaičių, sustabdytų dėl per ilgo darbo laiko komandų skaičių, „nestabilių“ sektorių skaičių, nepataisomų sektorių skaičių. Šios reikšmės atitinka šiuos atributus:

- ***Smart 5 Raw*** - perskirstytų sektorių skaičius;
- ***Smart 9 Raw*** - bendras valandinis disko darbo laikas;
- ***Smart 187 Raw*** - neištaisomų klaidų skaičius;
- ***Smart 188 Raw*** - sustabdytų dėl per ilgo darbo laiko komandų skaičius;
- ***Smart 194 Raw*** - disko temperatūra laipsniais Celcijaus;
- ***Smart 197 Raw*** - „nestabilių“ sektorių skaičius;
- ***Smart 198 Raw*** - nepataisomų sektorių skaičius;

4.1.1.3. Duomenų paruošimas

Eksperimentas atliktas naudojant individualius dienų failus, tačiau dėl didesnio duomenų kiekio faile bei bendros analizės buvo paruošti 4 papildomi CSV failai: kiekvienam mėnesiui po vieną ir viso ketvirčio bendras. Tam buvo parašytas BASH skriptas, kuriam per argumentą yra pateikiama „Backblaze“ CSV failų direktorija. Skriptas pavadintas „combine_csv.sh“; skriptas grąžina failus pavadinimais: „combine_all.csv“, „combine_01.csv“ (sausio mėnesiui), „combine_02.csv“ (vasario mėnesiui), „combine_03.csv“ (kovo mėnesiui). Skriptas yra pateiktas priedуose (Priedas nr. 2).

- Bendras CSV failas **„combine_all.csv“**:
  - Dydis: **2,5GB**;
  - Įrašų eilučių skaičius: **8 949 492**;
- Sausio mėnesio CSV failas **„combine_01.csv“**:
  - Dydis: **846MB**;
  - Įrašų eilučių skaičius: **3 039 306**;
- Vasario mėnesio CSV failas **„combine_02.csv“**:
  - Dydis: **782,3MB**;
  - Įrašų eilučių skaičius: **2 803 852**;
- Kovo mėnesio CSV failas **„combine_03.csv“**:
  - Dydis: **868,2MB**;
  - Įrašų eilučių skaičius: **3 106 334**.

4.2. Eksperimentui naudotos aplinkos aprašimas

Eksperimentas buvo atliekamas naudojant „Ubuntu (64-bit)“ operacinę sistemą įdiegtą virtualioje mašinoje „Oracle VM VirtualBox“. „MacroBase“ įdiegtas ir paruoštas darbui naudojantis „MacroBase“ dokumentaciją.

4.2.1. Virtuali mašina eksperimentui

„MacroBase“ pateikiami pavyzdžiai ir konfigūraciniai nurodymai yra pateikiami „Linux“ operacinėms sistemoms. Todėl eksperimentui atlikti buvo pasirinkta atviro kodo nemokama operacinė sistema „Ubuntu“. Dėl paprastumo buvo nuspręsta operacinei sistemai naudoti virtualią mašiną. Atviro kodo nemokama virtuali mašina „Oracle VM VirtualBox“ buvo pasirinkta šiai užduočiai. Galutinės eksperimentui naudotos sisteminės specifikacijos:

- „Ubuntu“ operacinės sistemos 64 bitų versija **„Ubuntu (64-bit)“**, versija: **16.04 LTS**;
- **„Oracle VM VirtualBox“**, versija: **5.2.12** r122591 (Qt5.6.2);
- Virtualus standusis diskas: **40GB**;
- Virtualiai mašinai skirta operatyvioji atmintis: **4GB**.

4.2.1.1. Virtualios mašinos paruošimas

Virtualios mašinos paruošimas darbui:

1. Iš „Oracle VM VirtualBox“ internetinės svetainės atsisiunčiamas naujausias „Windows 10“ (operacinė sistema į kuria diegiama virtuali mašina) operacinę sistemą palaikantis diegimo failas (`https://www.virtualbox.org/wiki/Downloads`).
2. Sekant sąrankos vedlio nurodymus įdiegiama „Oracle VM VirtualBox“ virtuali mašina.
3. Iš „Ubuntu“ internetinės svetainės atsisiunčiamas naujausios „Ubuntu (64-bit)“ operacinės sistemos ISO failas (`https://www.ubuntu.com/download/desktop`).
4. Atidarius „Oracle VM VirtualBox“ programinę įrangą pradedamas naujos operacinės sistemos pridėjimas spaudžiant ant mygtuko su tekstu „Nauja“.
5. Toliau rodomuose languose: pasirenkamas operacinės sistemos tipas - „Linux“; versija: „Ubuntu (64-bit)“; nurodomas operatyviosios atminties kiekis megabaitais: 4096MB; sukuriamas virtualus standusis diskas: 40GB.
6. „Oracle VM VirtualBox“ pagrindiniame lange pasirinkus naujai sukurtą virtualią mašiną spaudžiama ant mygtuko su tekstu „Paleisti“.
7. Atsiradusiame virtualios mašinos lange pasirenkamas prieš tai atsiųstas „Ubuntu (64-bit)“ ISO failas.
8. Pasirenkama „Install“ ir sekant diegimo vedlį, įdiegiama „Ubuntu (64-bit)“ operacinė sistema.

4.2.2. „MacroBase GUI“ analitinio įrankio su grafine naudotojo sąsaja paruošimas

Sekant „MacroBase“ dokumentacijoje nurodytus žingsnius įdiegiamas „MacroBase GUI“ [Inf17]:

1. Atidaromas „Ubuntu“ Terminalas;
2. Klonuojama projekto repozitorija:
   `git clone https://github.com/stanford-futuredata/macrobase.git`
3. Sukompiliuojamas „MacroBase“ ir paruošiamas darbui:
   `cd macrobase; mvn package`
4. Paleidžiamas „MacroBase“ serveris su grafinė sąsaja:
   `bin/frontend.sh`
5. Atidarius interneto naršyklę „MacroBase GUI“ pasiekiamas adresu:
   `http://localhost:8080`

Dirbant su CSV failas tai yra žingsniai, kurių užtenka parengti „MacroBase GUI“ darbui (Priedas nr. 3, pav. 1 ir 2).

4.2.2.1. „MacroBase GUI“ reikalingi papildomi paketai

Diegiant „MacroBase“ į naujai įdiegtą „švarią“ „Ubuntu“ operacinę sistemą, reikia įdiegti atitinkamus paketus. Visos reikalingos papildomos „Ubuntu“ terminalo komandos:

1. Versijavimo kontrolės sistemos „Git“ [Git18] diegimas:
   `sudo apt install git`
2. Java projektų valdymo ir diegimo priemonės „Maven“ [Pro18] diegimas:
   `sudo apt install maven`
3. Java JRE [Ora18a] diegimas:
   `sudo apt-get install default-jre`
4. Java JDK [Ora18a] diegimas:
   `sudo apt-get install default-jdk`

4.3. Eksperimento vykdymo eiga

Šiame poskyryje aprašoma eksperimento vykdymo eiga naudojant „Backblaze“ pateiktus 2018 metų pirmo ketvirčio duomenų rinkinius.

4.3.1. Eksperimento eiga su „Backblaze“ duomenimis

Su „Backblaze“ duomenų rinkiniais nagrinėjamas eksperimentas: Diskų, kurie turi didžiausia tikimybę sugesti artimoje ateityje aptikimas. „Backblaze“ naudoja SMART 5, 187, 188, 197 ir 198 (Priedas nr. 1) atributus diskų, kurie turi tikimybę sugesti numatymui [Kle16]. Eksperimento metu bus analizuojama, kokie diskų modeliai yra labiausiai linkę daryti kritines klaidas ir ar šie duomenys koruliuoja su „Backblaze“ fiksuojamomis klaidomis.

Šio eksperimento metu buvo paimtos pirmos 7 duomenų rinkinio dienos (2018-01-01 - 2018-01-07), iš „Backblaze“ įrašų sąrašo buvo pasirinktas vienas atributas - „Model“, nurodantis disko modelį, ir buvo ištirti tokie atvejai su visomis parinktomis dienomis:

1. Parinkta metrika: perskirstytų sektorių skaičiaus atributas (**smart 5 raw**);
2. Parinkta metrika: neištaisomų klaidų skaičiaus atributas (**smart 187 raw**);
3. Parinkta metrika: sustabdytų dėl per ilgo darbo laiko komandų skaičiaus atributas (**smart 188 raw**);
4. Parinkta metrika: „nestabilių“ sektorių skaičiaus atributas (**smart 197 raw**);
5. Parinkta metrika: nepataisomų sektorių skaičiaus atributas (**smart 198 raw**);
6. Parinkta metrika: „Backblaze“ klaidą nurodantis skaičiaus atributas (**failure**);

Kiekvienas rezultatas įsirašomas į atitinkamai pavadintą (pvz.: „failure_2018-01-01.txt“ ar „188_2018-01-05.txt“) tekstinį failą.

4.3.2. Eksperimento eiga su jungtiniais duomenimis

Su iš „Backblaze“ duomenų rinkinių sukurtais CSV failais buvo nagrinėjami eksperimentai:

1. Mėnesio dienos kuriomis diskai kaista labiausiai.
2. Darbas su dideliais CSV failais, jų krovimo laikas.

4.3.2.1. Eksperimentas: mėnesio dienos, kai diskų temperatūros yra neįprastai aukštos

Apjungtiems mėnesio duomenų CSV failams buvo parinktas datos atributas. Eksperimentui parinkta metrika buvo temperatūra. Atlikti tyrimai su visais trimis mėnesių failais atskirai. Šiuo eksperimentu nagrinėjama ar buvo dienų kuriomis buvo neįprastai aukšta temperatūra.

4.3.2.2. Eksperimentas: darbas su dideliais CSV failais

Apjungti mėnesių CSV failai ir bendras viso „Backblaze“ ketvirčio CSV failas yra žymaus dydžio failai (apie 800MB ir apie 2,5GB). Mėnesiniai failai susideda iš vidutiniškai 3 mln. įrašų eilučių; bendras failas susideda iš beveik 9 mln įrašų eilučių. Todėl šiuo eksperimentu tiriama ar parengta eksperimentinė aplinka yra pajėgi dirbti su tokio dydžio duomenų rinkiniais, ir kiek laiko su jais užtruks darbas. Eksperimentui parinktas disko modelis kaip atributas. Šiam eksperimentui

reikalinga paprasčiausia metrika, todėl metrikai buvo pasirinktas klaidos atributas, kuris gali įgyti tik dvi skirtingas reikšmes: „0“ arba „1“.

4.4. Eksperimento rezultatų analizė

Šiame poskyryje aprašyta atliktų eksperimentų su „Backblaze“ 2018 metų pirmo ketvirčio duomenų rinkiniais bei apjungtais pagal mėnesius ir visų duomenų bendrai viename rinkinyje rezultatų analizė.

4.4.1. Eksperimentų su „Backblaze“ duomenimis gauti rezultatai

Su „Backblaze“ duomenų rinkiniais buvo atliktas eksperimentas siekiant aptikti diskus, kurie turi didžiausią tikimybę sugesti. Eksperimento metu pasirinkta tirti pirmos duomenų rinkinio savaitės duomenis (2018-01-01 - 2018-01-07). Eksperimento metu surinkti duomenys randami prieduose (Priedas nr. 4 ir Priedas nr. 5).

Šio eksperimento metu buvo tiriami duomenys pagal atributus, kuriuos „Backblaze“ naudoja prie veikimo pabaigos artėjančių diskų atpažinimui. „MacroBase“ buvo užduota aiškinti metrikas pagal diskų modelio numerius, tokiu būdu gaunama informacija apie didžiausią tikimybę tam tikro modelio diskams, kiekvienos metrikos atžvilgiu, būti anomalijomis.

Pirmas tyrimas buvo atliktas su metrika perskirstytų sektorių skaičiumi. Naudojant šią metriką gautos vidutiniškai 708 išskirtys iš vidutiniškai 95143 įrašų (7 lentelė). Gautuose rezultatuose (1 lentelė) pastebima, jog anomalijas kelia tik „ST10000NM0086“ ir „Hitachi HDS722020ALA330“ diskai. „ST10000NM0086“ modelio diskai yra vidutiniškai turi 13,136537 kartų didesnę tikimybę būti anomalija, o „Hitachi HDS722020ALA330“ diskai net 15,956676 kartų. Net 106 disko modelio „ST10000NM0086“ įrašai yra perskirstytų sektorių anomalijos.

Antras tyrimas buvo atliktas naudojant neištaisomų klaidų skaičiaus metriką. Šio tyrimo metu bendras išskirčių skaičius yra mažesnis, nes ne visi diskai turi tyrimui naudotą metrikos atributą. Iš vidutiniškai 68018 įrašų 655 yra išskirtys (8 lentelė). Guotuose rezultatuos (2 lentelė) matomi trys skirtingi diskų modeliai: „ST4000DM000“, „ST31500541AS“ ir „ST320LT007“. „ST4000DM000“ modelio diskai sukėlė didžiausia kiekį anomalijų, bei turi didžiausią tikimybę būti anomalija - 11,09313357 kartų. Kol kas pasikartojimų tarp pirmo ir antro tyrimo nematome.

Trečio tyrimo metrika buvo sustabdymų skaičius dėl per ilgo komandos vykdymo laiko. Šito kaip ir antro tyrimo metu vidutinis įrašų skaičius yra mažesnis, nes ne visi įrašai turi tyrimui naudotą metrikos atributą. Vidutiniškai 2017 išskirčių ir 65847 neišskirčių (9 lentelė). Šitą anomaliją „MacroBase“ aiškino didesniu skaičiumi modelių nei praeitų tyrimų metu (3 lentelė). Diskai „ST4000DM000“, „ST9250315AS“, „ST31500541AS“ ir „ST320LT007“ buvo labiausiai išskirtos anomalijos. Šio tyrimo metu buvo pamatytas pirmas anomalijų persidengimas: diskai „ST4000DM000“ ir „ST31500541AS“ šio ir praeito tyrimo metu.

Ketvirtam tyrimui buvo parinkta „nestabilių“ (laukiančių perskirstymo) sektorių skaičiaus metrika. Šis tyrimas jau vyko su visai rinkinių įrašais - vidutiniškai 95143, iš kurių 334 buvo išskirtys (10 lentelė). Šiame tyrime matomas (4 lentelė) tas pats „ST4000DM000“ modelis bei pirmame tyrime matytas „Hitachi HDS722020ALA330“ modelis, taip pat du nauji „ST4000DM005“

ir „WDC WD1600AAJS“. Šį kartą „ST4000DM000“ turi tikimybę būti anomaliją ne tokią didelę kaip pirmo tyrimo metu, bet didesnę už trečio - 5,632494857 kartų „ST4000DM000“ modelio diskai turi didesnę tikimybę būti anomalijos.

Penktam ir paskutiniam eksperimento **tyrimui** naudojant SMART atributus buvo parinktas nepataisomų sektorių skaičiaus metrika. Iš vidutiniškai 95143 įrašų 301 buvo išskirtys ir 94842 neišskirtys (11 lentelė). Šio tyrimo metu (5 lentelė) buvo gauti net trys ankstesnių tyrimų metu gauti modeliai: „ST4000DM000“, „WDC WD1600AAJS“ ir „ST4000DM005“, bei vienas naudas „ST4000DM001“. Iš visų su SMART atributais atliktų tyrimų, galime pastebėti, kad „ST4000DM000“ diskai sukelia daugiausiai anomalijų ir yra nemaža tikimybė jiems sugesti.

Paskutinis atliktas tyrimas buvo su „Backblaze“ klaidos laukų, kuris įgyja reikšmę „1“, jei tai yra paskutinė disko veikimo diena ir jis yra pakeičiamas nauju. Šio tyrimo duomenys buvo rinkti iki šiol atliktų tyrimų informacijos patikrinimui. Taigi, iš 95165 įrašų - vidutiniškai 5 įrašai buvo išskirtys (12 lentelė). Kaip ir buvo spėta, dominuojantis anomalijų skaičiumi buvo „ST4000DM000“ disko modelis (6 lentelė). Kaip ir buvo matyti iš pirmų penkių tyrimų „ST4000DM000“ yra tirtosios savaitės labiausiai „Backblaze“ keičiamo modelio diskai ir didžiausią tikimybę gesti ateityje turintis modelis.

4.4.2. Eksperimentų su jungtiniais duomenimis gauti rezultatai

Su jungtiniais duomenų rinkiniais buvo atlikti du eksperimentai: išskirtinai aukštos diskų temperatūros dienos, ir darbas su dideliais CSV failais.

4.4.2.1. Dienų, kai diskų temperatūros yra išskirtinai aukštos eksperimento rezultatai

4.4.2.2. Darbo su dideliais CSV failais eksperimento rezultatai

Rezultatai ir išvados

Rezultatų ir išvadų dalyje turi būti aiškiai išdėstomi pagrindiniai darbo rezultatai (kažkas išanalizuota, kažkas sukurta, kažkas įdiegta) ir pateikiamos išvados (daromi nagrinėtų problemų sprendimo metodų palyginimai, teikiamos rekomendacijos, akcentuojamos naujovės).

Literatūra

[Ast16] Anthony Asta. Observability at twitter: technical overview, part i. 2016. URL: https://blog.twitter.com/engineering/en_us/a/2016/observability-at-twitter-technical-overview-part-i.html.

[Bac18] Backblaze. Hard drive data and stats. 2018. URL: https://www.backblaze.com/b2/hard-drive-test-data.html (tikrinta 2018-06-01).

[BGM+17] Peter Bailis, Edward Gan, Samuel Madden, Deepak Narayanan, Kexin Rong ir Sahaana Suri. Macrobase: prioritizing attention in fast data. *SIGMOD'17- Proceedings of the 2017 ACM International Conference on Management of Data*, p. 541–556, Chicago, Illinois, USA. Stanford Infolab ir Massachusetts Institute of Technology, ACM New York, 2017. ISBN: 978-1-4503-4197-4.

[BGR+17] Peter Bailis, Edward Gan, Kexin Rong ir Sahaana Suri. Prioritizing attention in fast data: principles and promise. *CIDR 2017, 8th Biennial Conference on Innovative Data Systems Research*. Stanford Infolab, 2017.

[Bun97] Peter Buneman. Semistructured data. *PODS '97 Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, p. 117–121, 1997. ISBN: 0-89791-910-6.

[Cer18] Cern. Processing: what to record? 2018. URL: https://home.cern/about/computing/processing-what-record (tikrinta 2018-06-05).

[DS16] Nedim Dedić ir Clare Stanier. Towards differentiating business intelligence, big data, data analytics and knowledge discovery. *Innovations in Enterprise Information Systems Management and Engineering*, p. 114–122, Hagenberg, Austria. Springer International Publishing, 2016.

[Eco10] The Economist. Data, data everywhere. 2010. URL: https://www.economist.com/node/15557443.

[EMC14] Dell EMC. The digital universe of opportunities: rich data and the increasing value of the internet of things. 2014. URL: http://www.emc.com/leadership/digital-universe/.

[Git18] Git. Git. 2018. URL: https://git-scm.com/ (tikrinta 2018-06-10).

[HA17] Makrufa Sh. Hajirahimova ir Aybeniz S. Aliyeva. About big data measurement methodologies and indicators. *International Journal of Modern Education and Computer Science*, 9:1–9, 2017.

[HC16] Vojtech Huser ir James J. Cimino. Impending challenges for the use of big data. *International Journal of Radiation Oncology*Biology*Physics*, 95:890–894, 2016.

[Hil14] Martin Hilbert. What is the content of the world's technologically mediated information and communication capacity: how much text, image, audio, and video? *The Information Society*, 30:127–143, 2014.

[Inf17] Standford InfoLab. Macrobase documentation. 2017. URL: https://macrobase.stanford.edu/docs/ (tikrinta 2018-06-10).

[Kle16] Andy Klein. What smart stats tell us about hard drives. 2016. URL: https://www.backblaze.com/blog/what-smart-stats-indicate-hard-drive-failures/.

[Lan01] Doug Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Tech. atask., META Group, 2001. URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[Loh13] Steve Lohr. The origins of 'big data': an etymological detective story. 2013. URL: https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/.

[Mar14] Bernard Marr. Big data: the 5 vs everyone must know. 2014. URL: https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know/.

[Mas98] John R. Mashey. Big data ... and the next wave of infrastres s. 1998. URL: http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf. Skaidrės.

[MCB+11] James Manyika, Michael Chui, Brad Brown, Richard Dobbs Jacques Bughin, Charles Roxburgh ir Angela Hung Byers. Big data: the next frontier for innovation, competition, and productivity. 2011. URL: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

[MCB+15] James Manyika, Michael Chui, Peter Bisson, Jonathan Woetzel, Richard Dobbs ir Jacques Bughin Dan Aharon. *McKinsey Global Institute: The internet of things: mapping the value beyond the hype*. 2015.

[MGG16] Andrea De Mauro, Marco Greco ir Michele Grimaldi. A formal definition of big data based on its essential features. 65, 2016.

[Ora18a] Oracle. Java se at a glance. 2018. URL: http://www.oracle.com/technetwork/java/javase/overview/index.html (tikrinta 2018-06-10).

[Ora18b] Oracle. What is big data? 2018. URL: https://www.oracle.com/big-data/guide/what-is-big-data.html (tikrinta 2018-06-05).

[PFT+15] Tuomas Pelkonen, Scott Franklin, Justin Teller, Paul Cavallaro, Qi Huang, Justin Meza ir Kaushik Veeraraghavan. Gorilla: a fast, scalable, in-memory time series database. *VLDB Endowment - Proceedings of the 41st International Conference on Very Large Data Bases*, p. 1816–1827, Kohala Coast, Hawaii. Facebook Inc., 2015.

[Pre13] Gil Press. A very short history of big data. 2013. URL: https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#7e969e1365a1.

[Pro18] Apache Maven Project. Apache maven. 2018. URL: https://maven.apache.org/ (tikrinta 2018-06-10).

[Sch16] Christie Schneider. The biggest data challenges that you might not even know you have. 2016. URL: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/.

[Sim71] Herbert A. Simon. *Designing organizations for an information-rich world*. *In Computers, communications, and the public interest*. Martin Greenberger, redaktorius. 1971, p. 37–72.

[Smi18] Kit Smith. 47 incredible facebook statistics and facts. 2018. URL: https://www.brandwatch.com/blog/47-facebook-statistics/ (tikrinta 2018-06-05).

[Smo14] Sandy Smolan. The human face of big data. Dan Oberle, redaktorius. Trumpametražis dokumentinis filmas, 2014.

[SMR12] Chris Snijders, Uwe Matzat ir Ulf-Dietrich Reips. „big data“: big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, 7:1–5, 2012. URL: http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf.

[Tec18a] Techopedia. Anomaly detection. 2018. URL: https://www.techopedia.com/definition/30297/anomaly-detection (tikrinta 2018-06-01).

[Tec18b] Techopedia. Big data. 2018. URL: https://www.techopedia.com/definition/27745/big-data (tikrinta 2018-06-05).

[Web18] Webopedia. Structured data. 2018. URL: https://www.webopedia.com/TERM/S/structured_data.html (tikrinta 2018-06-05).

[Wik18] Wikipedia. S.m.a.r.t. 2018. URL: https://en.wikipedia.org/wiki/S.M.A.R.T. (tikrinta 2018-06-10).

[Woo15] Alex Woodie. Kafka tops 1 trillion messages per day at linkedin. 2015. URL: https://www.datanami.com/2015/09/02/kafka-tops-1-trillion-messages-per-day-at-linkedin/.

Santrumpos

- **CSV** - „Failo formatas, sudarytas iš Kableliais atskirtų reikšmių“ angl. *Comma-Separated Values*.
- **BASH** - „Komandinių eilučių interpretatorius ir komandų kalba“.
- **JRE** - „JAVA vykdymo aplinka“ angl. *Java Runtime Environment*.
- **JDK** - „JAVA vystymo rinkinys“ angl. *Java Development Kit*.
- **S.M.A.R.T** arba **SMART** - „Save Stebinčios Analizuojančios ir Protokoluojančios Technologijos“ angl. *„Self-Monitoring, Analysis and Reporting Technology“*

Priedas nr. 1
„Backblaze“ duomenų SMART. atributų reikšmės

„Backblaze“ naudoja „S.M.A.R.T.“ kietųjų diskų protokolavimui. Tačiau „Backblaze“ naudoja tik dalį „S.M.A.R.T.“ atributų:

- ***Smart 1 raw, Smart 1 normalized*** - **Read Error Rate** - (nuo gamintojo priklausanti reikšmė) nuskaitymo klaidų dažnis (mažesnis geresnis).
- ***Smart 2 raw, Smart 2 normalized*** - **Throughput Performance** - bendras disko pralaidumo efektyvumas (didesnis geresnis).
- ***Smart 3 raw, Smart 3 normalized*** - **Spin-Up Time** - vidutinis laikas reikalingas pasiekti maksimalų disko sukimosi greitį (mažesnis geresnis).
- ***Smart 4 raw, Smart 4 normalized*** - **Start/Stop Count** - disko sukimo pradėjimų ir sustabdymų skaičius.
- ***Smart 5 raw, Smart 5 normalized*** - **Reallocated Sectors Count** - perskirstytų sektorių skaičius (mažesnis geresnis). Diskas, kurio nors vienas sektorius bent kartą yra perskirstytas, turi daug didesnę tikimybę sugesti.
- ***Smart 7 raw, Smart 7 normalized*** - **Seek Error Rate** - (nuo gamintojo priklausanti reikšmė) paieškos klaidų dažnis. Paieškos klaida atsiranda, kai įvyksta klaida mechaninėje pozicijos nustatymo sistemoje. Tai gali įvykti dėl įvairių priežasčių, pvz.: pažeidimo servo mechanizme, temperatūros sukelto plėtimosi, kt.
- ***Smart 8 raw, Smart 8 normalized*** - **Seek Time Performance** - vidutinis paieškos efektyvumas (didesnis geresnis). Krintanti reikšmė reiškia kylančias problemas mechaninėje posistemėje.
- ***Smart 9 raw, Smart 9 normalized*** - **Power-On Hours** - disko bendras valandų skaičius, kai diskas yra įjungtas.
- ***Smart 10 raw, Smart 10 normalized*** - **Spin Retry Count** - bandymų įsukti diską skaičius (mažesnis geresnis).
- ***Smart 11 raw, Smart 11 normalized*** - **Recalibration Retries or Calibration Retry Count** - pakartotinių kalibravimų skaičius (mažesnis geresnis).
- ***Smart 12 raw, Smart 12 normalized*** - **Power Cycle Count** - įjungimų ir išjungimų skaičius, žymi disko pilnų įjungimų ir išjungimų skaičių.
- ***Smart 13 raw, Smart 13 normalized*** - **Soft Read Error Rate** - neištaisytos nuskaitymo klaidos apie kurias yra pranešta operaciniai sistemai (mažesnis geresnis)
- ***Smart 15 raw, Smart 15 normalized*** - *informacijos apie šį atributą nerasta*.
- ***Smart 22 raw, Smart 22 normalized*** - **Current Helium Level** - (specifinis He8 diskams atributas) matuoja helio kiekį disko viduje.
- ***Smart 177 raw, Smart 177 normalized*** - **Wear Range Delta** - delta tarp labiausiai susidėvėjusių ir mažiausiai susidėvėjusių „Flash“ blokų.
- ***Smart 179 raw, Smart 179 normalized*** - **Used Reserved Block Count Total** - (Samsung diskų naudojamas atributas) - naudojamų rezervuotų blokų bendras skaičius.
- ***Smart 181 raw, Smart 181 normalized*** - **Program Fail Count Total or Non-4K Aligned Access Count** - bendras „Flash“ programos operacijų klaidų skaičius nuo disko naudojimo pradžios (mažesnis geresnis).
- ***Smart 182 raw, Smart 182 normalized*** - **Erase Fail Count** - (Samsung diskų naudojamas atributas) duomenų ištrynimų klaidų skaičius.

- ***Smart 183 raw, Smart 183 normalized*** - **SATA Downshift Error Count or Runtime Bad Block** - (Western digital, Samsung, Seagate atributas) - arba greičio disko sukimo greičio sumažėjimų skaičius arba bendras blokų su darbo metu aptiktomis, bet neištaisytomis klaidomis, skaičius (mažesnis geresnis).
- ***Smart 184 raw, Smart 184 normalized*** - **End-to-End error / IOEDC** - duomenų lygumo (angl. *parity*) klaidų skaičius (mažesnis geresnis).
- ***Smart 187 raw, Smart 187 normalized*** - **Reported Uncorrectable Errors** - skaičius klaidų, kurios negali būti ištaisytos naudojant techninę įrangą (mažesnis geresnis).
- ***Smart 188 raw, Smart 188 normalized*** - **Command Timeout** - skaičius atšauktų operacijų dėl pasibaigusio HDD laiko (angl. *timeout*) (mažesnis geresnis). Paprastai atributo reikšmė lygi „0“.
- ***Smart 189 raw, Smart 189 normalized*** - **High Fly Writes** - rašymo galvutės rašymo per aukštai nuo disko arba ne vietoje klaidų skaičius (mažesnis geresnis).
- ***Smart 190 raw, Smart 190 normalized*** - **Temperature Difference or Airflow Temperature** - reikšmė lygi $100 - temperatra$ °C, gamintojų naudojama nustatyti minimalų slenkstį, kuris atitinka maksimalią temperatūrą.
- ***Smart 191 raw, Smart 191 normalized*** - **G-sense Error Rate** - klaidų skaičius dėl išorinių smūgių ar vibracijos (mažesnis geresnis).
- ***Smart 192 raw, Smart 192 normalized*** - **Power-off Retract Count, Emergency Retract Cycle Count (Fujitsu), or Unsafe Shutdown Count** - nesaugių disko išjungimų skaičius (mažesnis geresnis).
- ***Smart 193 raw, Smart 193 normalized*** - **Load Cycle Count or Load/Unload Cycle Count (Fujitsu)** - skaičius užkrovimų ciklų skaičius (mažesnis geresnis).
- ***Smart 194 raw, Smart 194 normalized*** - **Temperature or Temperature Celsius** - nurodo disko temperatūrą (mažesnė geresnė).
- ***Smart 195 raw, Smart 195 normalized*** - **Hardware ECC Recovered** - (gamintojam specifinis atributas).
- ***Smart 196 raw, Smart 196 normalized*** - **Reallocation Event Count** - pertvarkymo operacijų skaičius (mažesnis geresnis).
- ***Smart 197 raw, Smart 197 normalized*** - **Current Pending Sector Count** - „nestabilių“ sektorių skaičius (mažesnis geresnis).
- ***Smart 198 raw, Smart 198 normalized*** - **(Offline) Uncorrectable Sector Count** - bendras neištaisytų klaidų skaičius skaitant arba rašant sektoriuje (mažesnis geresnis).
- ***Smart 199 raw, Smart 199 normalized*** - **UltraDMA CRC Error Count** - skaičius klaidų perduodant duomenis per laido sąsają (mažesnis geresnis).
- ***Smart 200 raw, Smart 200 normalized*** - **Multi-Zone Error Rate** - rastų klaidų skaičius rašant sektoriuje (mažesnis geresnis).
- ***Smart 201 raw, Smart 201 normalized*** - **Soft Read Error Rate or TA Counter Detected** - neištaisomų programinės įrangos skaitymo klaidų skaičius (mažesnis geresnis).
- ***Smart 222 raw, Smart 222 normalized*** - **Loaded Hours** - darbo laikas kraunant duomenis (judinant magnetinę galvutę).
- ***Smart 223 raw, Smart 223 normalized*** - **Load/Unload Retry Count** - galvutės pozicijos pakeitimų skaičius.
- ***Smart 224 raw, Smart 224 normalized*** - **Load Friction** - trinties sukeltų pasipriešinimų operuojant skaičius (mažesnis geresnis).

- ***Smart 225 raw, Smart 225 normalized*** - **Load/Unload Cycle Count** - bendras užkrovimų ciklų skaičius (mažesnis geresnis).
- ***Smart 226 raw, Smart 226 normalized*** - **Load 'In'-time** - bendras magnetinės galvutės krovimo laikas.
- ***Smart 235 raw, Smart 235 normalized*** - **Good Block Count AND System(Free) Block Count** - nurodo gerų blokų ir laisvų blokų skaičių.
- ***Smart 240 raw, Smart 240 normalized*** - Head Flying Hours or 'Transfer Error Rate' (Fujitsu) - bendras laikas praleista keičiant galvutės poziciją.
- ***Smart 241 raw, Smart 241 normalized*** - **Total LBAs Written** - bendras skaičius rašytų LBA.
- ***Smart 242 raw, Smart 242 normalized*** - **Total LBAs Read** - bendras skaičius skaitytų LBA.
- ***Smart 250 raw, Smart 250 normalized*** - **Read Error Retry Rate** - klaidų skaičius skaitant iš disko (mažesnis geresnis).
- ***Smart 251 raw, Smart 251 normalized*** - **Minimum Spares Remaining** - nurodo likusių nepanaudotų blokų skaičių ir procentinį jų kiekį.
- ***Smart 252 raw, Smart 252 normalized*** - **Newly Added Bad Flash Block** - bendras blogų „Flash" blokų skaičius nuo pirmo disko palaidimo.
- ***Smart 254 raw, Smart 254 normalized*** - **Free Fall Protection** - skaičius aptiktų „Laisvų Klaidos Įvykių" (angl. *„Free Fail Events"*.
- ***Smart 255 raw, Smart 255 normalized*** - *informacijos apie šį atributą nerasta*.

Priedas nr. 2

„Backblaze“ CSV failų apjungimo BASH skriptas

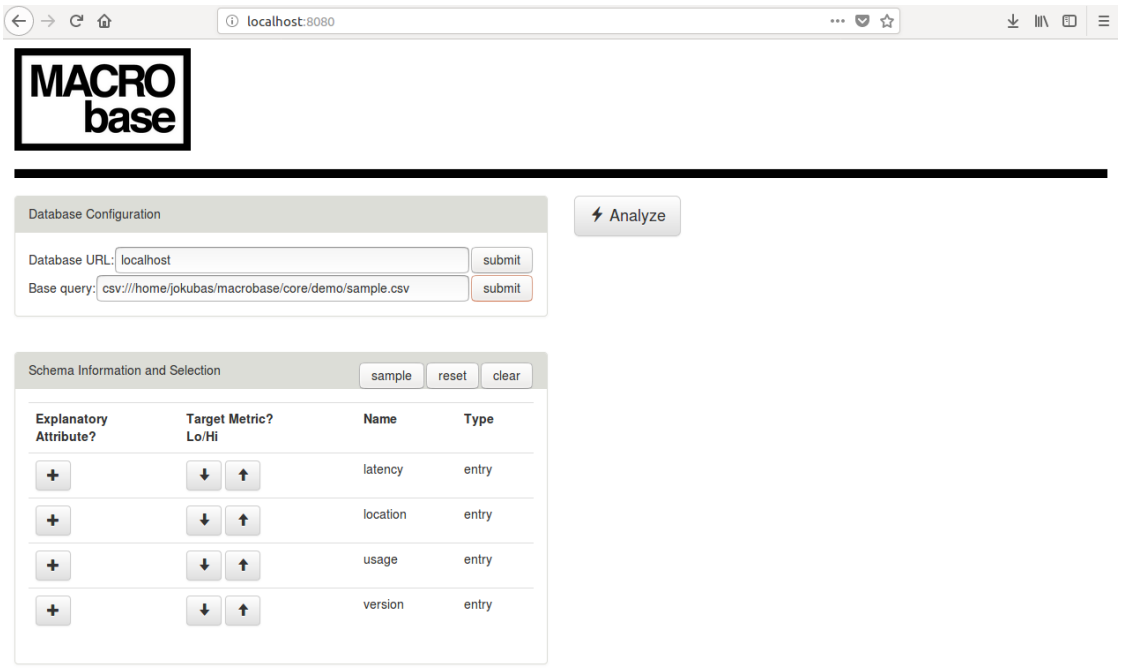
BASH skriptas „combine_csv.sh“ naudotas apjungti „Backblaze“ duomenis į vieną bendrą CSV failą ir 3 atskirus mėnesinius (sausiui, vasariui, kovui) CSV failus:

```
#!/bin/bash
directory=$1

head -1 $directory/2018-01-01.csv > combined_all.csv
head -1 $directory/2018-01-01.csv > combined_01.csv
head -1 $directory/2018-01-01.csv > combined_02.csv
head -1 $directory/2018-01-01.csv > combined_03.csv

for file_name in $(ls $directory/*.csv); do sed 1d $file_name >> combined_all.csv; done
for file_name in $(ls $directory/2018-01-*.csv); do sed 1d $file_name >> combined_01.csv;
done
for file_name in $(ls $directory/2018-02-*.csv); do sed 1d $file_name >> combined_02.csv;
done
for file_name in $(ls $directory/2018-03-*.csv); do sed 1d $file_name >> combined_03.csv;
done
```

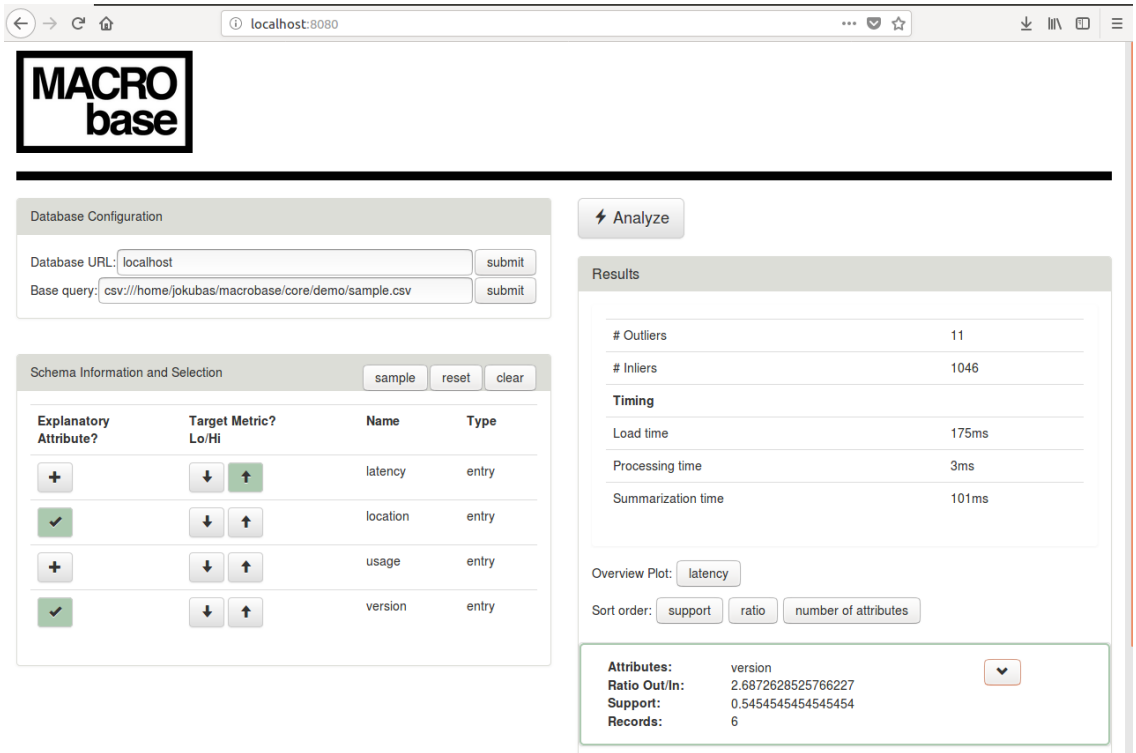
„Backblaze“ pateikiami duomenų failai yra pavadinti pagal įrašų rinkimo datą (pvz.: „2018-01-01.csv“).

Priedas nr. 3

„MacroBase GUI“ pagrindinis langas



1 pav. „MacroBase GUI“ pagrindinis langas



2 pav. „MacroBase GUI“ pagrindinis langas su analizės rezultatais

Priedas nr. 4

Diskų gedimų numatymo eksperimento rezultatai

1 lentelė. „Backblaze“ įrašų rinkinių rezultatai pagal SMART 5 metriką.

Data	Modelis	Santykis	Parama	Įrašai
2018-01-01	ST10000NM0086	13,276547	0,150142	106
	Hitachi HDS722020ALA330	22,368556	0,015581	11
2018-01-02	ST10000NM0086	13,184964	0,149086	106
	Hitachi HDS722020ALA330	22,237975	0,015471	11
2018-01-03	ST10000NM0086	13,415309	0,15113	107
	Hitachi HDS722020ALA330	22,358395	0,015537	11
2018-01-04	ST10000NM0086	13,407867	0,15113	107
	Hitachi HDS722020ALA330	22,346012	0,015537	11
2018-01-05	ST10000NM0086	13,414675	0,14986	107
	Hitachi HDS722020ALA330	22,385791	0,015406	11
2018-01-06	ST10000NM0086	11,119056	0,15461	109
2018-01-07	ST10000NM0086	14,137344	0,15461	109

2 lentelė. „Backblaze“ įrašų rinkinių rezultatai pagal SMART 187 metriką.

Data	Modelis	Santykis	Parama	Įrašai
2018-01-01	ST4000DM000	10,731441	0,907692	590
	ST31500541AS	3,92616	0,003077	2
	ST320LT007	9,593051	0,001538	1
2018-01-02	ST4000DM000	10,651343	0,90687	594
	ST31500541AS	3,893089	0,003053	3
	ST320LT007	9,521604	0,001527	1
2018-01-03	ST4000DM000	10,950332	0,90937	592
	ST31500541AS	3,932132	0,003072	2
	ST320LT007	9,60599	0,001536	1
2018-01-04	ST4000DM000	11,202331	0,911315	596
	ST31500541AS	3,910297	0,003058	2
	ST320LT007	9,558811	0,001529	1
2018-01-05	ST4000DM000	11,293594	0,909786	595
	ST31500541AS	3,984939	0,003058	2
	ST320LT007	9,720288	0,001529	1
2018-01-06	ST4000DM000	11,382906	0,907436	598
	ST31500541AS	4,047396	0,003035	2
	ST320LT007	9,855501	0,001517	1
2018-01-07	ST4000DM000	11,439988	0,907576	599
	ST31500541AS	4,04827	0,00303	2
	ST320LT007	9,857412	0,001515	1

3 lentelė. „Backblaze“ įrašų rinkinių rezultatai pagal SMART 188 metriką.

Data	Modelis	Santykis	Parama	Įrašai
2018-01-01	ST4000DM000	2,944763	0,731426	1585
	ST9250315AS	30,817261	0,020305	44
	ST31500541AS	6,581815	0,004153	9
	ST320LT007	27,368309	0,003692	8
2018-01-02	ST4000DM000	2,934509	0,730306	1576
	ST9250315AS	30,961801	0,019926	43
	ST31500541AS	6,622138	0,004171	9
	ST320LT007	27,523551	0,003707	8
2018-01-03	ST4000DM000	2,918827	0,729755	1577
	ST9250315AS	30,987409	0,020361	44
	ST31500541AS	6,620992	0,004165	9
	ST320LT007	27,519139	0,003702	8
2018-01-04	ST4000DM000	2,94161	0,731494	1591
	ST9250315AS	30,798674	0,02023	44
	ST31500541AS	6,577933	0,004138	9
	ST320LT007	27,353363	0,003678	8
2018-01-05	ST4000DM000	3,019693	0,731282	1592
	ST9250315AS	31,204788	0,020211	44
	ST31500541AS	6,672341	0,004134	9
	ST320LT007	27,71684	0,003675	8
2018-01-06	ST4000DM000	3,210627	0,736015	1592
	ST9250315AS	32,021745	0,020342	44
	ST31500541AS	6,861249	0,004161	9
	ST320LT007	28,444145	0,003699	8
2018-01-07	ST4000DM000	3,063261	0,726277	1592
	ST9250315AS	31,642066	0,020073	44
	ST31500541AS	6,774701	0,004106	9
	ST320LT007	28,11094	0,00365	8

4 lentelė. „Backblaze“ įrašų rinkinių rezultatai pagal SMART 197 metriką.

Data	Modelis	Santykis	Parama	Įrašai
2018-01-01	ST4000DM000	5,581127	0,745455	246
	ST4000DM005	2,785014	0,00303	1
	Hitachi HDS722020ALA330	2,416696	0,00303	1
	WDC WD1600AAJS	5,179099	0,00303	1
2018-01-02	ST4000DM000	5,615303	0,746224	247
	ST4000DM005	2,777484	0,003021	1
	Hitachi HDS722020ALA330	2,409746	0,003021	1
	WDC WD1600AAJS	4,826323	0,003021	1
2018-01-03	ST4000DM000	5,718268	0,75	249
	ST4000DM005	2,689523	0,003012	1
	Hitachi HDS722020ALA330	2,400746	0,003012	1
	WDC WD1600AAJS	4,81239	0,003012	1
2018-01-04	ST4000DM000	5,679619	0,749245	248
	ST4000DM005	2,625275	0,003021	1
	Hitachi HDS722020ALA330	2,411657	0,003021	1
	WDC WD1600AAJS	4,829281	0,003021	1
2018-01-05	ST4000DM000	5,72532	0,747748	249
	ST4000DM005	2,644973	0,003003	1
	Hitachi HDS722020ALA330	2,430448	0,003003	1
	WDC WD1600AAJS	4,858351	0,003003	1
2018-01-06	ST4000DM000	5,666128	0,741935	253
	ST4000DM005	2,597019	0,002933	1
	WDC WD1600AAJS	4,787528	0,002933	1
2018-01-07	ST4000DM000	5,441699	0,733918	251
	ST4000DM005	2,589356	0,002924	1
	WDC WD1600AAJS	4,776212	0,002924	1

5 lentelė. „Backblaze“ įrašų rinkinių rezultatai pagal SMART 198 metriką.

Data	Modelis	Santykis	Parama	Įrašai
2018-01-01	ST4000DM000	9,2748	0,828283	246
	WDC WD1600AAJS	5,976668	0,003367	1
	ST4000DM005	3,316613	0,003367	1
	ST4000DM001	2,271584	0,013468	4
2018-01-02	ST4000DM000	9,330921	0,828859	247
	WDC WD1600AAJS	5,582143	0,003356	1
	ST4000DM005	3,306457	0,003356	1
	ST4000DM001	2,265228	0,013423	4
2018-01-03	ST4000DM000	9,575818	0,832776	249
	WDC WD1600AAJS	5,564131	0,003344	1
	ST4000DM005	3,207011	0,003344	1
	ST4000DM001	2,257359	0,013378	4
2018-01-04	ST4000DM000	9,708536	0,835017	248
	WDC WD1600AAJS	5,610976	0,003367	1
	ST4000DM005	3,154697	0,003367	1
	ST4000DM001	2,320308	0,013468	4
2018-01-05	ST4000DM000	9,704559	0,832776	249
	WDC WD1600AAJS	5,638116	0,003344	1
	ST4000DM005	3,17309	0,003344	1
	ST4000DM001	2,332064	0,013378	4
2018-01-06	ST4000DM000	9,141099	0,821429	253
	ST4000DM001	2,27726	0,012987	4
	WDC WD1600AAJS	5,514558	0,003247	1
	ST4000DM005	3,089422	0,003247	1
2018-01-07	ST4000DM000	8,762776	0,814935	251
	ST4000DM001	2,281408	0,012987	4
	WDC WD1600AAJS	5,52402	0,003247	1
	ST4000DM005	3,095831	0,003247	1

6 lentelė. „Backblaze“ įrašų rinkinių rezultatai pagal „Failure“ metriką.

Data	Modelis	Santykis	Parama	Įrašai
2018-01-01	ST4000DM000	2,48423	0,666667	2
	HGST HUH728080ALE600	42,566018	0,333333	1
2018-01-03	ST4000DM000	8,846895	0,833333	5
2018-01-04	ST4000DM000	Infinity	1	4
2018-01-05	ST12000NM0007	3,758622	0,333333	2
2018-01-06	ST4000DM000	4,888753	0,75	3
2018-01-07	ST4000DM000	Infinity	1	2

Priedas nr. 5

Diskų gedimų numatymo eksperimento rezultatų statistikos

7 lentelė. „Backblaze“ įrašų rinkinių rezultatu statistikos pagal SMART 5 metriką.

Data	...-01	...-02	...-03	...-04	...-05	...-06	...-07
Išskirtys	706	711	708	708	714	705	705
Neišskirtys	93480	93612	93707	93656	94627	95921	96043
Krovimo Laikas	1186ms	1390ms	1576ms	1316ms	1890ms	1558ms	1654ms
Vykdymo Laikas	180ms	220ms	245ms	238ms	418ms	255ms	214ms
Apibendrinimo Laikas	40ms	75ms	43ms	82ms	43ms	43ms	46ms

8 lentelė. „Backblaze“ įrašų rinkinių rezultatu statistikos pagal SMART 187 metriką.

Data	...-01	...-02	...-03	...-04	...-05	...-06	...-07
Išskirtys	650	655	651	654	654	659	660
Neišskirtys	66406	66497	66584	66614	67563	68881	68997
Krovimo Laikas	1506ms	2050ms	1747ms	1412ms	1220ms	1903ms	1633ms
Vykdymo Laikas	196ms	314ms	200ms	164ms	160ms	288ms	229ms
Apibendrinimo Laikas	30ms	32ms	28ms	22ms	24ms	37ms	27ms

9 lentelė. „Backblaze“ įrašų rinkinių rezultatu statistikos pagal SMART 188 metriką.

Data	...-01	...-02	...-03	...-04	...-05	...-06	...-07
Išskirtys	2167	2158	2161	2175	2177	2163	2192
Neišskirtys	64889	64994	65074	65093	66040	67377	67465
Krovimo Laikas	2071ms	1622ms	1747ms	1996ms	1936ms	1683ms	1514ms
Vykdymo Laikas	427ms	288ms	226ms	199ms	222ms	293ms	234ms
Apibendrinimo Laikas	38ms	46ms	34ms	33ms	39ms	48ms	34ms

10 lentelė. „Backblaze“ įrašų rinkinių rezultatu statistikos pagal SMART 197 metriką.

Data	...-01	...-02	...-03	...-04	...-05	...-06	...-07
Išskirtys	330	331	332	331	333	341	342
Neišskirtys	93856	93992	94083	94033	95008	96285	96406
Krovimo Laikas	2441ms	1758ms	2111ms	2337ms	1471ms	1565ms	1584ms
Vykdymo Laikas	255ms	296ms	252ms	249ms	217ms	242ms	276ms
Apibendrinimo Laikas	85ms	47ms	42ms	48ms	65ms	41ms	44ms

11 lentelė. „Backblaze“ įrašų rinkinių rezultatu statistikos pagal SMART 198 metriką.

Data	...-01	...-02	...-03	...-04	...-05	...-06	...-07
Išskirtys	297	298	299	297	299	308	308
Neišskirtys	93889	94025	94116	94067	95042	96318	96440
Krovimo Laikas	1572ms	1911ms	2244ms	3161ms	1597ms	1602ms	1704ms
Vykdymo Laikas	354ms	254ms	288ms	279ms	229ms	214ms	229ms
Apibendrinimo Laikas	43ms	45ms	56ms	41ms	51ms	42ms	44ms

12 lentelė. „Backblaze“ įrašų rinkinių rezultatu statistikos pagal „Failure“ metriką.

Data	...-01	...-02	...-03	...-04	...-05	...-06	...-07
Išskirtys	3	8	6	4	6	4	2
Neišskirtys	94183	94315	94409	94360	95522	96622	96746
Krovimo Laikas	1643ms	2100ms	1730ms	2270ms	1702ms	2157ms	2258ms
Vykdymo Laikas	252ms	355ms	223ms	221ms	239ms	255ms	250ms
Apibendrinimo Laikas	46ms	57ms	38ms	41ms	41ms	47ms	49ms