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

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 („Kableliu Atskirtos Reikšmės“ (angl. *„Comma-Separated Values“*)) 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], „Save Stebinčios Analizuojančios ir Protokoluojančios Technologijos“ (angl. *„Self-Monitoring, Analysis and Reporting Technology“* dažniausiai sutrumpintai vadinami „SMART“), laukai:

- ***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 laukas. „0“, jei kietasis diskas dirba korektiškai; „1“, jei tai buvo paskutinė diena, kai diskas buvo naudojamas prieš sugesdamas.
- ***SMART*** laukai - 100 stulpelių duomenų iš kurių 50 yra neapdorotų ir 50 normalizuotų duomenų stulpelių (laukų reikšmės yra aprašytos prieduose (Priedas nr. 1)).

4.1.1.2. Duomenų paruošimas

Eksperimentui 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 Shell“ 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 prieduose (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 grafine 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, 1).

4.2.2.1. „MacroBase GUI“ reikalingi papildomi paketai

Diegiant „MacroBase“ į naują „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

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).

[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

Sąvokų apibrėžimai ir santrumpų sąrašas sudaromas tada, kai darbo tekste vartojami specialūs paaiškinimo reikalaujantys terminai ir rečiau sutinkamos santrumpos.

Priedas nr. 1
„Backblaze“ duomenų S.M.A.R.T. laukų paskirtys

„Backblaze“ naudoja „S.M.A.R.T.“ technologija kietųjų diskų protokolavimui. Tačiau „Backblaze“ naudoja ne visus „S.M.A.R.T.“ laukų:

- ***Smart 1 raw, Smart 1 normalized*** - (nuo gamintojo priklausanti reikšmė) nuskaitymo klaidų dažnis (mažesnis geresnis).
- ***Smart 2 raw, Smart 2 normalized*** - bendras disko pralaidumo efektyvumas (didesnis geresnis).
- ***Smart 3 raw, Smart 3 normalized*** - vidutinis laikas reikalingas pasiekti maksimalų disko sukimosi greitį (mažesnis geresnis).
- ***Smart 4 raw, Smart 4 normalized*** - disko sukimo pradėjimų ir sustabdymų skaičius.
- ***Smart 5 raw, Smart 5 normalized*** - 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*** - (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*** - vidutinis paieškos efektyvumas (didesnis geresnis). Krintanti reikšmė reiškia kylančias problemas mechaninėje posistemėje.
- ***Smart 9 raw, Smart 9 normalized*** - disko bendras valandų skaičius, kai diskas yra įjungtas.
- ***Smart 10 raw, Smart 10 normalized*** - bandymų įsukti diską skaičius (mažesnis geresnis).
- ***Smart 11 raw, Smart 11 normalized*** - pakartotinių kalibravimų skaičius (mažesnis geresnis).
- ***Smart 12 raw, Smart 12 normalized*** - įjungimų ir išjungimų skaičius, žymi disko pilnų įjungimų ir išjungimų skaičių
- ***Smart 13 raw, Smart 13 normalized*** -
- ***Smart 15 raw, Smart 15 normalized*** -
- ***Smart 22 raw, Smart 22 normalized*** -
- ***Smart 177 raw, Smart 177 normalized*** -
- ***Smart 179 raw, Smart 179 normalized*** -
- ***Smart 181 raw, Smart 181 normalized*** -
- ***Smart 182 raw, Smart 182 normalized*** -
- ***Smart 183 raw, Smart 183 normalized*** -
- ***Smart 184 raw, Smart 184 normalized*** -
- ***Smart 187 raw, Smart 187 normalized*** -
- ***Smart 188 raw, Smart 188 normalized*** -
- ***Smart 189 raw, Smart 189 normalized*** -
- ***Smart 190 raw, Smart 190 normalized*** -
- ***Smart 191 raw, Smart 191 normalized*** -
- ***Smart 192 raw, Smart 192 normalized*** -

- ***Smart 193 raw, Smart 193 normalized*** -
- ***Smart 194 raw, Smart 194 normalized*** -
- ***Smart 195 raw, Smart 195 normalized*** -
- ***Smart 196 raw, Smart 196 normalized*** -
- ***Smart 197 raw, Smart 197 normalized*** -
- ***Smart 198 raw, Smart 198 normalized*** -
- ***Smart 199 raw, Smart 199 normalized*** -
- ***Smart 200 raw, Smart 200 normalized*** -
- ***Smart 201 raw, Smart 201 normalized*** -
- ***Smart 222 raw, Smart 222 normalized*** -
- ***Smart 223 raw, Smart 223 normalized*** -
- ***Smart 224 raw, Smart 224 normalized*** -
- ***Smart 225 raw, Smart 225 normalized*** -
- ***Smart 226 raw, Smart 226 normalized*** -
- ***Smart 235 raw, Smart 235 normalized*** -
- ***Smart 240 raw, Smart 240 normalized*** -
- ***Smart 241 raw, Smart 241 normalized*** -
- ***Smart 242 raw, Smart 242 normalized*** -
- ***Smart 250 raw, Smart 250 normalized*** -
- ***Smart 251 raw, Smart 251 normalized*** -
- ***Smart 252 raw, Smart 252 normalized*** -
- ***Smart 254 raw, Smart 254 normalized*** -
- ***Smart 255 raw, Smart 255 normalized*** -

Priedas nr. 2

„Backblaze“ CSV failų apjungimo „Bash Shell“ skriptas

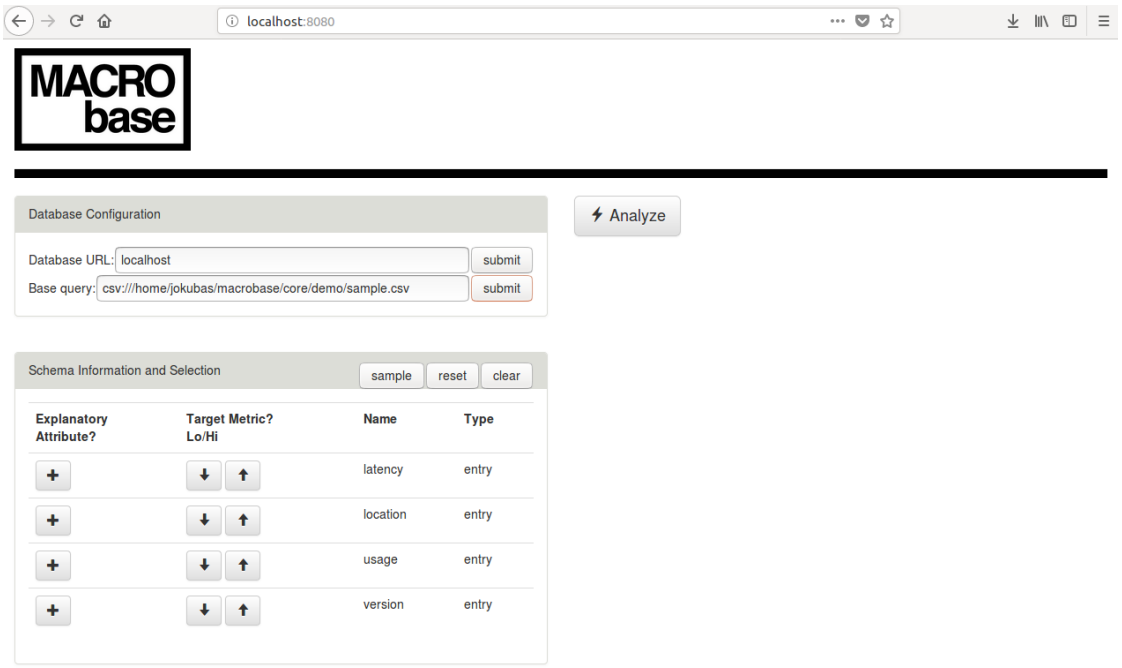
„Bash Shell“ skriptas „combine_csv.sh“ naudotas apjungti „Backblaze“ duomenis į vieną bendrą CSV failą ir 3 atskirus mėnesinius (sausiui, vasariui, kovui) CSV failus:

```
#!/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
```

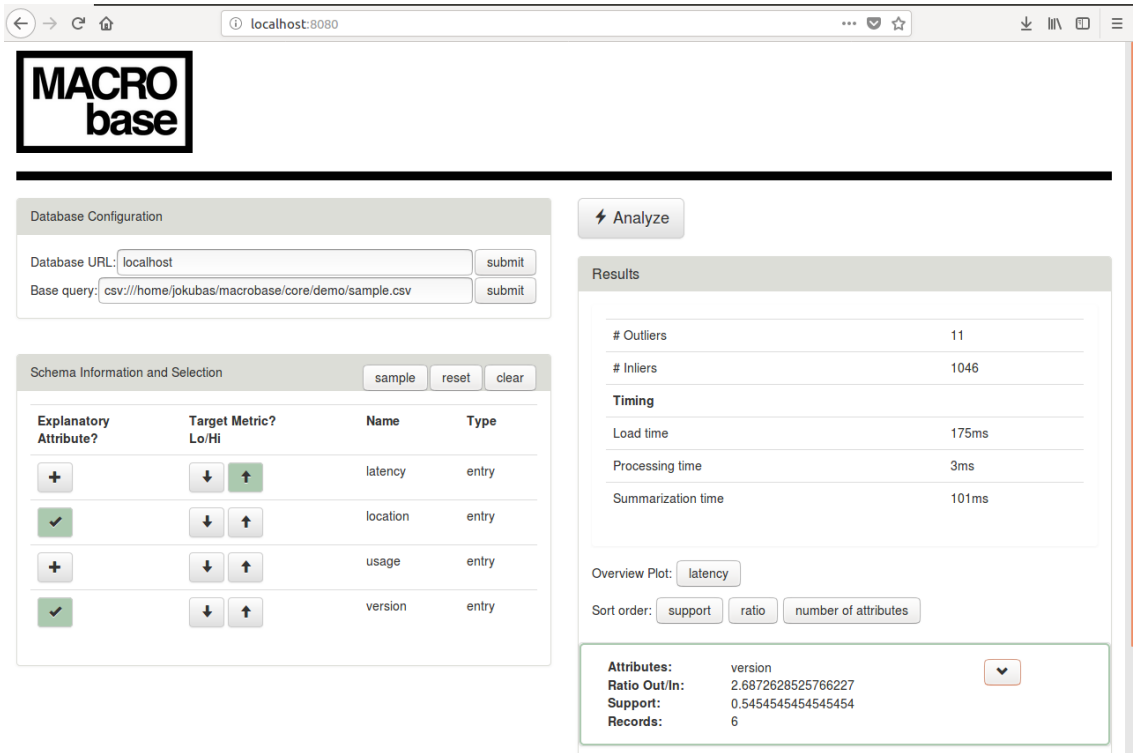
„Backblaze“ pateikiami duomenų failai yra pavadinti pagal įrašų rinkimo datą (pvz.: „2018-01-01.csv“).

Priedas nr. 3

„MacroBase GUI“ pagrindinis langas



1 pav. „MacroBase GUI“ pagrindinis langas



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