VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS INSTITUTAS
PROGRAMŲ SISTEMŲ BAKALAURO STUDIJŲ PROGRAMA

Debesų kompiuterijos paslaugų, skirtų duomenų gavybai, palyginamoji analizė

The Comparative Analysis of Cloud Computing Services Dedicated to Data Mining

Bakalauro baigiamasis darbas

Atliko:	Michal Stankevič	(parašas)
Darbo vadovas:	j. asist. Raimundas Savukynas	(parašas)
Darbo recenzentas:	lekt. dr. Edvinas Pakalnickas	(parašas)

Vilnius – 2021

Santrauka

Duomenyse, kurie šiandien renkami daugumoje žmonių veiklų, glūdi daug naudingos informacijos. Tą informaciją įgalina išgauti duomenų gavyba. Esant dideliam debesų kompiuterijos paslaugų, skirtų duomenų gavybai, pasirinkimui, svarbu apibrėžti tam tikras gaires, pagal kurias galima pasirinkti optimalų variantą. Darbe palyginamos duomenų gavybos paslaugos „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“. Pirma, apibrėžiami duomenų gavybos paslaugų palyginimo aspektai - funkcionalumas, vartotojo sąsaja, integralumas ir našumas. Antra, aprašomas apgaulingų kreditinių kortelių transakcijų aptikimo modelio programinis įgyvendinimas. Modelis verifikuojamas naudojant minėtas paslaugas, tuo pačiu vertinama vartotojo sąsaja bei funkcionalumas iš praktinės perspektyvos. Trečia, atliekama paslaugų našumo analizė matuojant užklausų apdorojimo greitį. Apibendrinant šių žingsnių metu gautus rezultatus nustatyta, kad vartotojo patogumo aspektu geriausia naudoti „Microsoft Azure HDInsight“ paslaugą, našumo aspektu - „Google Cloud Dataproc“, o atsižvelgus į visus aspektus ir kainos kriterijų: optimalu naudoti „Google Cloud Dataproc“.

Raktiniai žodžiai: duomenų gavyba, duomenų gavybos paslaugos, palyginamoji analizė

Summary

The huge amount of data collected on a variety of human activities contains a lot of useful information that is difficult for humans to analyze and comprehend. Data mining enables extraction of knowledge from such data. There is a lot of services that provide data mining functionality and therefore it is important to define guidelines for choosing the optimal service. This paper demonstrates a comparative analysis of data mining tools, specifically: ”Microsoft Azure HDInsight”, ”Google Cloud Dataproc” and ”Amazon Elastic MapReduce”. Firstly, the comparison aspects of data mining services are defined: functionality, user interface, integrity and performance. Secondly, the data mining model that predicts fraudulent credit cards transactions is developed. While the model is being further validated using the aforementioned services, user experience and functionality of services are evaluated. Lastly, the performance analysis of the aforementioned services is conducted. Summarized results of intermediate steps suggest that regarding the user experience aspect the best choice is ”Microsoft Azure HDInsight”, regarding the performance aspect - ”Google Cloud Dataproc”. When all the aspects and price are considered then the optimal choice is ”Google Cloud Dataproc”.

Keywords: data mining, data mining services, comparative analysis

TURINYS

Įvadas

Didėjantis informacinių technologijų (toliau IT) panaudojimas įvairiose žmonių veiklos srityse suponuoja intensyvų duomenų kaupimą bei poreikį pasinaudoti sukaupta informacija priimant sprendimus. Duomenų bazėse sukaupti neapdoroti duomenys nėra palankūs interpretavimui ir supratimui, tačiau šiuolaikinės technologijos, konkrečiai - duomenų gavyba - įgalina įžvelgti tam tikrus dėsnius, tendencijas, modelius - žinias, glūdinčias didžiuosiuose duomenyse. Šiame darbe analizuojama keletas paslaugų, skirtų duomenų gavybai - „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“.

Generuojamų duomenų kiekis auga itin sparčiai, prognozuojama globali duomenų apimtis (angl. *The Global Datasphere*) 2025 metais siekia apie 175 zetabaitus[RGR18]. Savarankiška žmogaus atliekama duomenų analizė nepraktiška laiko ir resursų atžvilgiu[FPS96a]. Informacijos pritaikymas suteikia privalumų ir savo ruožtu lemia verslo konkurencingumą. Technologija, įgalinanti atrasti žinias, paslėptas didžiuosiuose duomenyse vadinama duomenų gavyba. Sukauptuose milžiniškuose duomenyse slypi itin vertinga informacija. Įrašai apie apsipirkimus leidža prekybos įmonėms prognozuoti pirkėjų elgesį. Bankai, rinkdami vartotojų transakcijų duomenis, prognozuoja, kuriems klientams rizikinga duoti paskolas; analizuojant surinktus duomenis galima aptikti anomalijas ir finansines machinacijas[MA12]. Gydymų duomenys gali būti naudingi sveikatos ekspertams vertinant, kokie gydymo metodai efektyvesni[KT]. Pagal paieškos sistemoje naudotas užklausas vartotojui rodomos tinkintos reklamos. Telekomunikacijų sektoriuje analizuojant vartotojų pokalbių, paslaugų, sunaudotų duomenų informaciją galima parinkti optimalų pasiūlymą klientui, aptikti tinklo operacijų riktus bei identifikuoti įtartinus, apgavikiškus vartotojus[Jos13].

Duomenų gavyba yra sudėtingas ir netrivialus procesas, kurio prasmė - pritaikius specifinius algoritmus iš duomenų gauti duomenų modelius. Duomenų gavyba įgalina žinių išgavimą iš milžiniškų duomenų, šį procesą konceptualiai sudaro tokie žingsniai: duomenų surinkimas, išankstinis duomenų apdorojimas, transformacija, modeliavimas ir vertinimas[FPS96b]. Duomenų gavyba yra tarpdisciplininė sritis, apjungianti statistiką, mašininį mokymąsį bei duomenų bazių teoriją.

Didėjanti paklausa, poreikis versle pritaikyti informacinių technologijų IT sprendimus, leidžiančius apdoroti duomenis implikuoja didėjančią atitinkamų sprendimų pasiūlą, todėl egzistuoja platus debesų kompiuterijos paslaugų, skirtų duomenų gavybai asortimentas. Nėra paprasta pasirinkti tinkamą paslaugą žinant verslo poreikius, nes egzistuoja daugybė antraeilių veiksnių, pavyzdžiui, nefunkciniai reikalavimai, kurių ignoravimas ženkliai sumažina naudojamos paslaugos produktyvumą. Norint pasirinkti optimalią duomenų gavybos paslaugą, kad pritaikius ir integravus ją į veiklos srityje naudojamą sistemą, būtų patenkinti vartotojų lūkesčiai ir būtų kuriama pridėtinė vertė, pravartu remtis apibrėžtais aspektais.

Darbo tikslas - palyginti debesų kompiuterijos paslaugas, skirtas duomenų gavybai - „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“ bei realizuoti duomenų gavybos procesą panaudojant šias paslaugas.

Darbo uždaviniai:

1. Išanalizuoti duomenų gavybos procesą, išskiriant jo sudedamuosius žingsnius.
2. Nustatyti aspektus, kuriais remiantis bus sugretinamos duomenų gavybos paslaugos.
3. Išnagrinėti „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“ dokumentaciją ir palyginti šias paslaugas pagal nustatytus aspektus.
4. Realizuoti duomenų gavybos procesą atsižvelgiant į apibrėžtus žingsnius, panaudojant duomenų gavybos paslaugas „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“ ir įvertinti šių paslaugų gebėjimą atlikti užduotį.

Pirmame skyriuje pristatomas mokslinis įdirbis nagrinėjama tema, apžvelgiami susiję darbai ir apibendrinama literatūra. Antrame skyriuje analizuojamas duomenų gavybos konceptas, paaiškinami jo sudedamieji žingsniai ir pristatomi iššūkiai, susiję su šio proceso praktiniu pritaikymu. Taip pat apibrėžiami duomenų gavybos paslaugų palyginimo aspektai. Trečiame skyriuje paslaugos „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“ yra analizuojamos keliais pjūviais. Pirma, nagrinėjama dokumentacija, aprašomos minėtų paslaugų techninės galimybės ir kitos ypatybės. Antra, atliekamas praktinis šių paslaugų vertinimas realizuojant duomenų gavybos procesą. Trečia, vertinamas efektyvumas matuojant duomenų apdorojimo laiką. Galiausiai atliekama palyginamoji analizė pagal apibrėžtus aspektus.

1. Literatūros analizė

Per pastaruosius dvidešimt metų įvyko duomenų gavybos proceso sąveikos su vartotoju raida. Pradedant nuo sudėtingų programinės įrangos sprendimų, prieinamų tik aukštos kvalifikacijos programuotojams, naudojamų tik kaip darbalaukio programos, turinčias neintuityvią grafinę sąsają, iki debesų kompiuterijoje prieinamų paslaugų, kurios leidžia itin lanksčiai atlikti duomenų inžinierijos procesus. Atitinkamai kito reikalavimai šiai programinės įrangos grupei, todėl moksliniai tyrimai atspindi pokytį vertinimo aspektuose. Toliau chronologine tvarka apžvelgiami tyrimai apie duomenų gavybos, analizės ar inžinierijos įrankius.

King, Elder ir kiti palygino keturiolika įrankių, skirtų duomenų gavybai[KEG+98]. Darbo rezultate apibrėžta daugiau nei 20 aspektų, sugrupuotų į penkias kategorijas: pajėgumas (angl. *capability*), panaudojamumas (angl. *usability*), sąveika (angl. *interoperability*), lankstumas ir tikslumas. Įrankių vertinime dalyvavo trys žmonių grupės, pagal kompetenciją dirbti su duomenų gavyba; testuota naudojant keturis duomenų rinkinius. Praktinėje dalyje sprendžiami klasifikacijos uždaviniai naudojant sprendimų medį, taisyklių indukciją, neuroninius tinklus.

Giraud-Carrier ir Povel apibrėžė duomenų gavybos įrankių vertinimo aspektų taksonomiją[GP03], aukščiausiame lygyje yra šios kategorijos: verslo tikslas (kokią problemą sprendžia), modelio tipas (kokius algoritmus naudojant iš duomenų gaunamas modelis), duomenų gavybos proceso metu naudojamų algoritmų prieinamumas, vartotojo sąsaja, sistemos reikalavimai ir įrankio gamintojo informacija. Tyrimo autoriai atkreipė dėmesį į tai, kad apibrėžtais aspektais vertinami įrankiai būtų naudojami išimtinai priimant verslo sprendimus.

Wahbeh, Al-Radaideh ir kiti atliko tyrimą[WAA+], kuriame palygino „Waikato Environment for Knowledge Analysis“, „RapidMiner“, „Clementine“, „Rosetta“, „Intelligent Miner“ - programinę įrangą, aprėpiančią metodus ir algoritmus duomenų gavybai. Atlikta išsami duomenų gavybos proceso tikslumo analizė naudojant klasifikacijos algoritmus - „Naive Bayes“, „One Rule“, „Zero rule“, k-artimiausias kaimynas. Tyrimas parodė iššūkius, susijusius su klasifikavimo įgyvendinimu. Pavyzdžiui, diskretūs duomenys neapdorojami kai kurių klasifikavimo algoritmų. Tyrime vertinta, kokie įrankiai siūlo konkrečius klasifikavimo algoritmus ir jų tikslumą.

Solanki palygino duomenų gavybos įrankius „Waikato Environment for Knowledge Analysis“, „Konstanz Information Miner“, „Tanagra“[Sol13] ir įvertino juos pagal duomenų formatų palaikymos galimybes, grafinį rezultato pavaizdavimą, duomenų analizės ir apdorojimo galimybes ir išskyrė specifinius įrankių trūkumus, dažniausiai susijusius su duomenų apdorojimu ir duomenų kiekiu.

Verma ir Dhawan pasiūlė karkasą duomenų gavybos įrankių pasirinkimui. Vertinimo aspektai suskirstyti į keturias grupes: atlikimas, funkcionalumas, naudojamumas ir palaikymas[VD14]. Vertinimo modelis iš dalies tinka darbo objekto analizei, bet reiketų atmesti nereikalingas charakteristikas, tokias kaip: platformų pasirinkimas, kadangi debesų kompiuterijos paslaugos kliento neapriboja technologiniu - platformų ir operacinių sistemų - atžvilgiu.

Muntean palygino lyderiaujančių debesų kompiuterijos tiekėjų paslaugas, susijusias su verslo analitika[MUN15]. Autorius atkreipė dėmesį į tokius aspektus: infrastruktūros užtikrinimas

debesų kompiuterija, duomenų saugyklos (angl. *data warehouse*) ir „ETL“ proceso (angl. *Extract, Transform, Load*) galimybės, duomenų radimo (angl. *data discovery*) galimybės, savitarna, mišrių šaltinių integravimas (angl. *connectors for hybrid sources*).

Vaughan ir Chen palygino „Google Trends“ ir „Baidu Index“[VC15]. Šios paslaugos skirtos analizuoti ir išgauti žinias iš užklausų, įvedamų ieškojimo sistemose. Išskirtos šios funkcionalumo ypatybės (angl. *features*): duomenų kategorizavimas pagal regionus, pagal dalykines kategorijas (pavyzdžiui, apibrėžus kategoriją - technologijos - užklausa „Apple“ rodys rezultatus, susijusius su šia technologijos įmone, išvengiant dviprasmybės), maksimalus palyginamų sąvokų skaičius, vidutinė paieškos apimtis, atitikties metodai (angl. *term matching*).

Poggi, Montero ir Carrera[PMC17] išnagrinėjo debesų kompiuterijos paslaugas, susijusias su duomenų analize ir gavyba. Darbe minima „ALOJA“ platforma - Barselonos superkompiuterių centro (angl. *Barcelona Supercomputing Center*) studentų iniciatyva sukurta tam, kad galima būtų tyrinėti ir įvertinti paslaugas, siūlančias programinės ar geležinės įrangos architektūrinius sprendimus, skirtus realizuoti didžiųjų duomenų apdorojimą[Cen].

Dakic, Stefanovic ir kiti palygino atviro kodo duomenų gavybos paslaugas[DSS+17] pagal tokias aspektų grupes: bendros charakteristikos (architektūros tipas, prieinamos programavimo kalbos, operacinės sistemos reikalavimai), duomenų valdymas (duomenų formatai, duomenų šaltinių tipai, duomenų dydis). Taip pat šiame darbe atlikti klasifikacijos uždaviniai ir įvertinti įrankių tikslumai - gautų rezultatų atitikimas teisingiems rezultatams.

Apibendrinant susijusius darbus, gaunama duomenų gavybos paslaugų vertinimo aspektų aibė: funkcionalumas, vartotojo sąsaja, integralumas. Trečiame skyriuje, aiškinant darbo metodus, šie aspektai bus detaliau aprašyti.

2. Duomenų gavybos paslaugų palyginimo aspektai

Šiame skyriuje detaliau nagrinėjamas duomenų gavybos konceptas, paaiškinami jo sudedamieji žingsniai. Taip pat patikslinami susijusių darbų apibendrinime apibrėžti duomenų gavybos paslaugų palyginimo aspektai.

2.1. Duomenų gavybos proceso apibrėžimas

Šiandien dėl padidėjusio žmonių mobilumo, pakitusios žmogaus su kompiuteriu sąveikos, paspartintos išaugusiu interneto panaudojimu, diskutuoti apie duomenų gavybą paminint debesų kompiuteriją būtų netikslu, todėl svarbu paaiškinti šį konceptą. Debesų kompiuterija tai technologija, leidžianti užtikrinti IT paslaugas internetu[AFG+09]. Alternatyviai apibrėžiama kaip paradigma, žadanti neribotą kompiuterinių išteklių kiekį[NSC+16]. Debesų kompiuterija įgalina verslą naudojant IT privalumus susitelkti ties esminiais veiklos aspektais, nesirūpinant dėl IT infrastruktūros palaikymo ir užtikrinimo. Kokybiškas, pajėgus ir našus IT veikimas reikalauja sudėtingų sisteminių įgyvendinimų, kurie nesusiję su pagrindine verslo veikla, todėl poreikis integruoti IT į verslo aplinką priverčia rinktis: verslas pats rūpinsis technine ir programine įrangomis, ar naudosis debesų kompiuterijos paslaugomis. Verslo ir akademinėje bendruomenėje sutarta, kad egzistuoja trijų lygių paslaugos, užtikrinamos debesų kompiuterija: infrastruktūros lygio, platformos lygio ir programinės įrangos lygio.

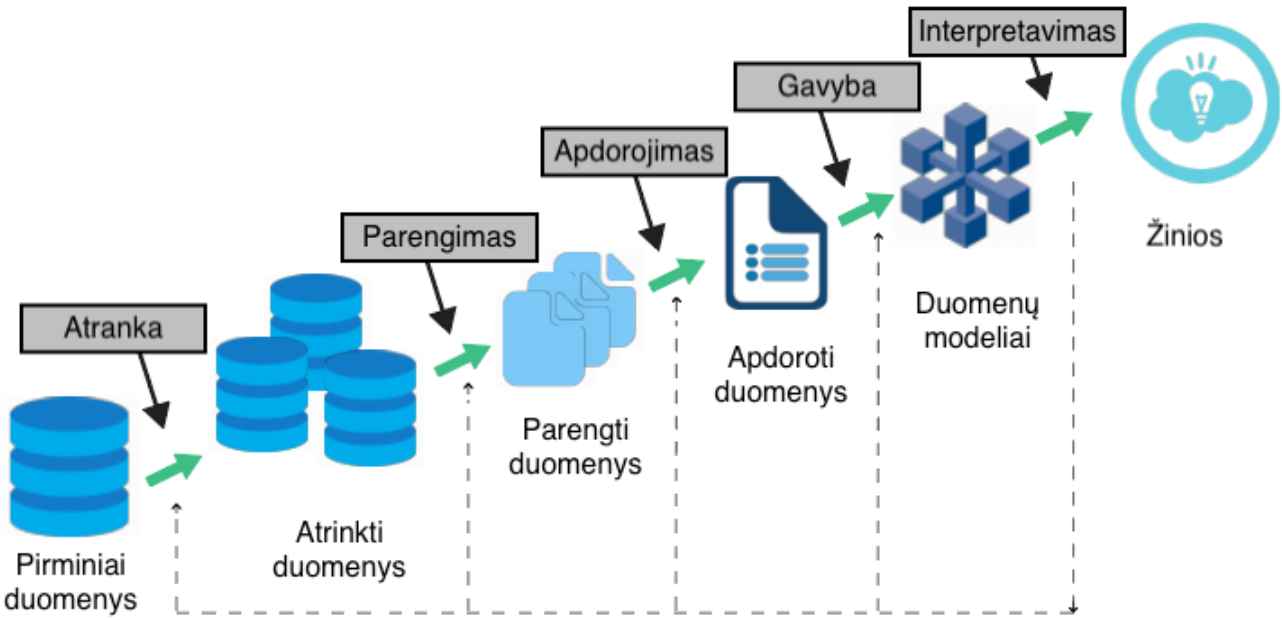
Duomenų gavyba neatsiejama nuo „didžiųjų duomenų“ (angl. *Big Data*). Sąvoka „didieji duomenys“ naudojama apibrėžti duomenų aibes, kurių sudėtingumas ir didelė apimtis reikalauja unikalių ir pažengusių technologinių sprendimų šiems duomenims talpinti, valdyti, apdoroti, pavaizduoti[CCS12]. 2001-ais buvo pasiūlytas naujas duomenų valdymo modelis, turėsiantis patenkinti augančius komercinių veiklų poreikius, susijusius su duomenų valdymu, atsiradusius dėl populiarėjančios elektroninės komercijos[Lan01a]. Tuomet buvo išskirti trys esminiai iššūkiai, sukuriantys rimtų problemų tradiciniams duomenų apdorojimo metodams - apimtis, sparta, įvairovė (angl. *Volume, Velocity, Variety*).

- Apimtis. Duomenų apimtis priklauso nuo vartotojų skaičiaus ir duomenų išsamumo. Šie dalykai neišvengiamai didėja/gerėja bet kurios įmonės raidoje. Elektroninė komercija leido padidinti dalykinės transakcijos duomenų „pralaidumą“ - padaroma daugiau transakcijų, ir kiekvienoje transakcijoje surenkama daugiau duomenų[Lan01b]. „Google“ teigimu[Goob], apie kiekvieną „Google“ paslaugų vartotoją renkama tokia informacija: peržiūrėti vaizdo įrašai; skelbimai, ant kurių spustelėta; vietovė; aplankytos svetainės bei to vartotojo pateikta ar sukurta informacija: vardas, pavardė, gimimo data; el. laiškai; komentarai sistemoje „YouTube“ ir taip toliau. „Google“ paslaugos „Gmail“ paskyrų skaičius perkopė pusantro milijardo.
- Sparta. Elektroninė komercija drastiškai padidino tempą, kuriuo vykdomos transakcijos[Lan01c]. Teigiama, kad sparta yra svarbesnė už duomenų apimtį, mat informacija, gaunama realiu laiku leidžia įmonėms būti judresnemis už konkurentus[MB12]. Tai suponuoja reikalavimą duomenų valdymo technologijoms užtikrinti atitinkamą spartą

atliekant susijusias duomenų operacijas[Lan01c].

- Įvairovė. Technologiniai pasiekimai duomenų kaupimui įgalina naudoti tiek struktūruotus, tiek pusiau struktūruotus, tiek nestruktūruotus duomenų formatus[MB14].

Duomenų gavyba tai procesas, kurio metu gaunama neišreikštinė, anksčiau nežinoma ir potencialiai naudinga informacija[LS01]. Svarbu paminėti, kad terminas „duomenų gavyba“ vartojamas netiksliai (angl. *misnomer*), iš tikrųjų jis galėtų būti pakeistas „žinių gavyba“, bet tolesniame darbe, jeigu neminima kitaip, šie terminai naudojami ekvivalentiškai. Duomenų gavyba įgalina aptikti netikėtus ir vertingus duomenų modelius, kurių pritaikymas naujiems duomenims gali duoti teigiamus ir verifikuojamus rezultatus. Konceptualiai duomenų gavybos procesas pavaizduotas paveikslėlyje 1.



1 pav. Žinių gavybos proceso žingsniai

Duomenų kokybę lemia tokie faktoriai: tikslumas, išsamumas, neprieštaringumas[VH00]. Natūralu, kad pirminiai duomenys netenkina šių faktorių, todėl svarbu prieš realizuojant pagrindinį duomenų gavybos žingsnį, paruošti duomenis, kad jie būtų geresnės kokybės. Toliau įvardijami ir detaliau paaiškinami duomenų gavybos proceso žingsniai.

1. Duomenų atranka. Pirminiuose duomenyse yra nepilnų įrašų ir anomalijų, kurie gali sukelti netikslumų atliekant tolesnius žingsnius ir iškraipyti rezultatus, todėl svarbu užtikrinti duomenų išsamumą. Nepilnų įrašų problema sprendžiama dviem būdais: pilnai ištrinamas įrašas arba trūkstamų atribūtų reikšmės užpildomos labiausiai tikėtinomis reikšmėmis, pavyzdžiui, kitų įrašų netuščių atributų moda, mediana, vidurkis arba numatytoji reikšmė. Anomalijų šalinime naudojamas reikšmių glodinimas grupuojant įrašus - (angl. *data binning*).
2. Duomenų parengimas. Šiame žingsnyje užtikrinamas duomenų integralumas, kas yra itin svarbu kuomet duomenys yra heterogeniški. Pavyzdžiui, duomenys surenkami iš

duomenų bazių, failų ir kitų duomenų saugojimo formatų, tuomet dažnu atveju skiriasi esybių ir atributų pavadinimo standartai, sukeliant metaduomenų perteklių. Todėl duomenų integralumas leidžia pagerinti tikslumą ir paspartinti duomenų gavybos procesą.

3. Duomenų apdorojimas. Šiame žingsnyje atliekama redukcija ir transformacija. Redukcija įgalina gauti reikšmingus analizei duomenis - mažesne apimtimi, tačiau kartu pakankamai reprezentatyvius. Yra keli būdai redukuoti duomenis. Pirma, mažinant dimensijų skaičių: nereikalingi atributai pašalinami visoje duomenų aibėje. Turint dalykinės srities žinias apibrėžiami (ne)reikalingi atributai arba naudojant pažingsninio parinkimo algoritmus (angl. *Step-wise selection*) sugrupuojami atributai į gerus ir blogus. Antra, mažinant įrašų skaičių. Ištrinami besidubliuojantys įrašai rankiniu būdu arba mažinama duomenų aibė panaudojant regresijos, klasterizacijos, kubinės duomenų agregacijos (angl. *data cube aggregation*), pavyzdžių atrankos ar duomenų suspaudimo metodus. Transformacijos metu duomenys įgauna optimalų pavidalą atlikti duomenų gavybos procedūrą. Įprastai transformacija susideda iš toliau įvardijamų žingsnių: glodinimas, agregavimas, diskretizacija, generalizacija bei normalizavimas. Glodinimas - procesas kurio metu šalinamas duomenų triukšmas. Duomenų triukšmas apsunkina duomenų analitikams įžvelgti tendencijas ir duomenų struktūrą. Glodinime naudojami klasterizacija ir regresija. Agregavimo proceso metu atliekamos įvairios operacijos su duomenimis, kurio rezultate gaunami apibendrinti duomenys. Diskretizacija - tai tolydžių duomenų suskaidymas į mažesnes dalis, pavyzdžiui, intervalus. Generalizacijos proceso metu detalūs atributai gali būtų keičiami bendresniais, tai priklauso nuo koncepto hierarchijos. Normalizavimo metu daromos duomenų projekcijos į tam tikrą aibę. Taigi atlikus redukciją ir transformaciją gaunami optimalūs, duomenų gavybai paruošti duomenys.
4. Modeliavimas. Pagrindinis duomenų gavybos žingsnis, kuriame panaudojant įvairias modeliavimo metodikas iš duomenų nustatomos arba aptinkamos duomenų struktūra bei tendencijos. Priklausomai nuo dalykinės srities tikslų, parenkami atitinkami modeliavimo metodai, sukuriami testai modelio kokybės ir validumo patikrinimui. Vienas iš primityvių būdų validuoti modelį - padalinti paruoštus testinius duomenis į dvi dalis, vieną naudoti modelio apmokymui, kitą - testavimui. Parinkti metodai paleidžiami ant paruoštų duomenų, toliau gautas rezultatas - duomenų modelis - validuojamas naudojant testus ir vertinamas duomenų analitikų. Modeliavime naudojami šie metodai: klasterizacija, klasifikacija, regresija, asociacijų taisyklės.
5. Rezultatų interpretavimas. Atrasti duomenų modeliai vertinami jau dalykinės srities kontekste, ar atitinka reikalavimus. Jeigu rezultatas netenkina, procesas kartojamas iš naujo. Priklausomai nuo modelio teisingumo, rezultatas gali būti panaudotas naujoje iteracijoje kaip įvestis. Jei rezultatas tenkina, modelis diegiamas naudoti su realiais duomenimis. Jei yra poreikis, gauti rezultatai suprantamu ir aiškiu pavidalu pristatomi suinteresuotiems asmenims - interaktyvios diagramos, ataskaitos, lentelės.

2.2. Darbo metodai

Šiame poskyryje paaiškinami debesų kompiuterijos paslaugų, skirtų duomenų gavybai, palyginamosios analizės sudedamieji žingsniai. Pirma, išsamiai analizei reikia išnagrinėti dokumentaciją, kad būtų aiškios, bent teoriniame lygmenyje, paslaugų funkcinės bei techninės galimybės. Taip pat reikia apibrėžti palyginimo aspektus, kuriais remiantis galima sugretinti paslaugas. Idealiu atveju aspektas turėtų įgalinti išrikiuoti paslaugas pagal įvertinimą, kitaip tariant atitiktų kriterijaus apibrėžimą. Sudėtinga būtų išryškinti daug kriterijų paslaugoms, todėl šiame etape tikslas yra apibrėžti pakankamus aspektus, kurie leistų sudaryti kuo išsamesnį vaizdą apie analizuojamas paslaugas. Antra, įgyvendinti duomenų gavybos procesą. Autoriaus manymu šis žingsnis įgalina atlitki praktinį paslaugų vertinimą. Taip pat parodoma, kaip duomenų gavybos paslaugos gali būti naudingos sprendžiant realiame pasaulyje kylančias problemas. Trečia, atlikti techninę paslaugų analizę, remiantis, bet neapsiribojant, etaloninio bandymo karkaso „TPCx-BB“ gairėmis[TPC; TPC21].

2.3. Palyginimo aspektai

Remiantis literatūros analize pasirinkti tokie duomenų gavybos paslaugų palyginimo aspektai: funkcionalumas, našumas, vartotojo sąsaja, integralumas. Toliau detaliau paaiškinami kiekvienas iš aspektų:

1. Funkcionalumas. Vertinant paslaugą šiuo aspektu nustatoma, kokios funkcinės galimybės pasiūlomos. Svarbu atkreipti dėmesį į tai, neužtenka vien įvardyti, kokias ypatybes gali pasiūlyti duomenų gavybos paslauga, turi būti įvertinta siūlomų galimybių kokybė. Tai reikalauja papildomų įžvalgų tiek iš teorinės, tiek iš praktinės pusės. Nagrinėjant funkcionalumą, svarbu atsižvelgti į tokius faktorius: duomenų importavimo ir eksportavimo galimybės, duomenų gavybos kodo keitimo galimybės, modelio validavimas, duomenų transformacijos ir paruošimo galimybės, duomenų gavybos metodų prieinamumas, modelio ir rezultatų vizualizavimo galimybės, programavimo aplinkų pasirinkimas, duomenų operacijų interaktyvaus analitinio apdorojimo režime (angl. *OLAP*) prieinamumas.
2. Vartotojo sąsaja. Vertinami paslaugos naudojimo patogumo ir efektyvumo faktoriai: grafinės sąsajos sprendimai, vizualizuoto programavimo galimybės, ar yra vartotojo instrukcijų, tiesioginė pagalba, mokymosi kreivė, klaidų sekimas ir žurnalizavimas, tekstinės sąsajos galimybės.
3. Integralumas. Vertinama, kokios duomenų bazių valdymo sistemos ir programavimo aplinkos integruotinos su duomenų gavybos paslauga bei kiti duomenų analitikos procesus įgalinantys įrankiai.

2.4. Duomenų gavybos proceso įgyvendinimo eiga

Įgyvendinant duomenų gavybos procesą pademonstruojama, kaip duomenų gavybos paslaugos naudojamos realiame pasaulyje ir kokia gaunama nauda. Tuo pat metu empiriniu metodu įvertinamos duomenų gavybos paslaugos funkcinės, techninės galimybės ir kokybė kitais aspektais.

Šiame poskyryje paaiškinama, kaip bus įgyvendinamas duomenų gavybos procesas, jo tikslas, sudedamieji žingsniai, įvesties duomenys ir kokio rezultato tikimasi.

Užduotis: duomenų gavybos būdu sukurti modelį, atpažįstantį apgaulingas kreditinių kortelių transakcijas. Pradiniai duomenys[KU18; PCJ+15] paimti iš duomenų inžinierijos ir mašininio mokymosi bendruomenės „Kaggle“ išteklių, prieinamų registruotiems vartotojams.

Duomenų rinkinio aprašymas. Duomenys pateikia informaciją apie atsiskaitymus Europos šalių vartotojų kreditinėmis kortelėmis 2013-ųjų metų rugsėjo mėnesį. Kiekviena duomenų eilutė interpretuojama kaip atskira transakcija. Duomenų rinkinys nesubalansuotas, nes apgaulingų transakcijų yra tik 0,172% (492 iš 284 807 transakcijų) .Visų atributų duomenų tipas yra skaičius. Daugelis atributų yra principinės komponentų analizės (angl. *Principal Component Analysis*) rezultatas, jų pavadinimai neturi aiškios prasmės (*V1, V2, ..., V28*). Metaduomenyse taip pat nepaaiškinta šių atributų prasmė. Likę atributai, kurie nėra principinės komponentų analizės rezultatas tai „laikas“, „kiekis“ ir „klasė“. Atributas „laikas“ nurodo, kiek sekundžių praėjo nuo pirmos transakcijos, atributas „kiekis“ nurodo, kiek pinigų turėtų būti pervesta. Atributas „klasė“ gali įgyti dvi reikšmes - 0, jeigu transakcija validi arba 1, jeigu transakcija buvo apgaulinga. Toliau pažingsniui pristatomas duomenų gavybos proceso programinis įgyvendinimas, komentuojant atliekamus veiksmus ir pateikiant programinį kodą. Naudojama *R* programavimo kalba.

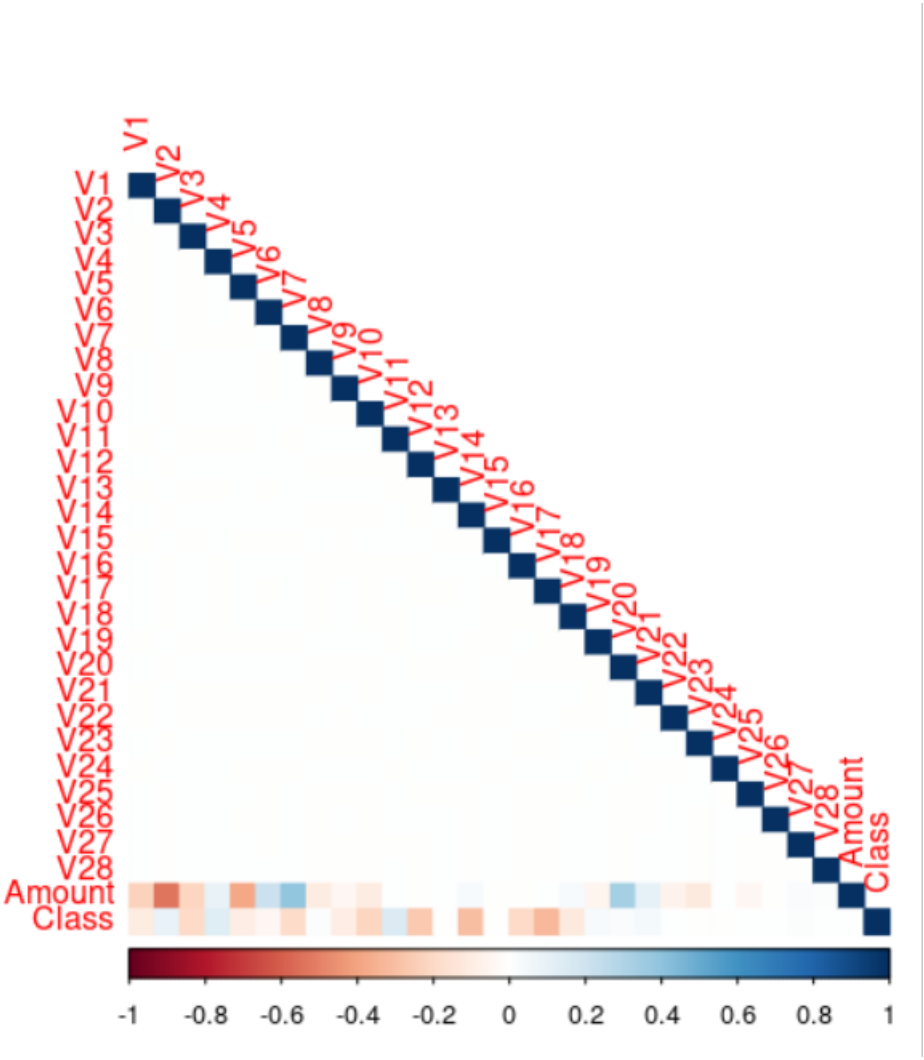
1. Atliekamas duomenų nuskaitymas. Pirminiai duomenys pateikti viename „CSV“ formato faile („cctransactions.csv“).

   ```
   duomenys <- fread("data/cc_transactions.csv")
   ```

2. Duomenų atranka. Standartiškai, šio proceso metu tikrinamas duomenų anomalijų buvimas ir ar yra trūkstamų duomenų atributų. Tačiau šio uždavinio kontekste anomalijos gali būti svarbios, nes jos gali koreliuoti su apgaulingomis transakcijomis, todėl anomalijų tikrinimo žingsnis ignoruojamas. Trūkumų tikrinimas atliekamas taip: kiekvienam atributui (stulpeliui) pritaikoma funkcija, kuri sumuoja, kiek yra nepasiekiamų (angl. *not available*) langelių.

   ```
   apply(duomenys, 2, function(x) sum(is.na(x)))
   ```

   Rezultate kiekvienam atributui pritaikius šią funkciją reikšmė yra 0, taigi nėra trūkstamų duomenų.
3. Duomenų parengimas. Kadangi pirminiai duomenys pateikti viename faile ir nuskaityti, o nėra kitų heterogeniškų duomenų šaltinių, nereikia atlikti papildomų operacijų. Žingsnis praleidžiamas.
4. Duomenų apdorojimas. Kadangi dauguma atributų neturi aiškios semantinės prasmės, nėra galimybės panaudoti dalykinės srities žinių nereikalingų atributų eliminavimui. Tokiu atveju sudaroma atributų koreliacijos lentelė.



2 pav. Atributų koreliacijos grafikas

Aukščiau pateiktame grafike matoma, kaip atributai tarpusavyje koreliuoja - negatyvi koreliacija žymima raudona spalva, pozityvi koreliacija - mėlyna. Pašalinami atributai, kurie nekoreliuoja su „klasės“ atributu, šiuo atveju tai yra atributai *V13, V15, V24, V26, V27, V28*.

```
optimizuoti_duomenys <- duomenys[-c(13, 15, 24, 26:28)]
```

5. Modelio kūrimas. Iš visų duomenų modelio apmokymui pasirenkama 70%, likę 30% skirti modelio testavimui.

```
apmokymo_indeksas <- createDataPartition( optimizuoti_duomenys$Class,
                                          times = 1, p = 0.7,
                                          list = FALSE)
apmokymas_x <- optimizuoti_duomenys[apmokymo_indeksas]
testavimas_x <- optimizuoti_duomenys[!apmokymo_indeksas]
testavimas_y <- optimizuoti_duomenys$Class[-apmokymo_indeksas]
```

Dėl didelės duomenų disproporcijos „klasės“ atributo atžvilgiu, naudojama „SMOTE“ metodika[CBH+02]. Atliekama duomenų pavyzdžių atranka (angl. *resampling*).

```
kontrol <- trainControl(method = "cv",number = 10, verboseIter = T,
classProbs = T,  sampling = "smote", summaryFunction = twoClassSummary,
savePredictions = T)
```

Panaudojant atsitiktinių medžių rinkinio (angl. *random forest*) metodą, apmokomas modelis. Taip pat sklandžiam darbui su duomenimis, dėl specifinių *R* programavimo kalbos priežasčių atliekama atributų pervadinimo operacija.

```
apmokymas_x_amr <- apmokymas_x
apmokymas_x_amr$Class <- as.factor(apmokymas_x_amr$Class)
levels(apmokymas_x_amr$Class) <- make.names(c(0, 1))
modelis <- train(Class ~ ., data = apmokymas_x_amr, method = "rf",
trControl = kontrol, metric="ROC")
```

Apmokytas modelis testuojamas ant testinės duomenų rinkinio dalies.

```
rezultatas <- predict(modelis, testavimas_x, type = "PROB")
```

6. Rezultatų interpretavimas. Turėdami rezultatų matricą, galima apibendrinti modelio charakteristikas.

```
rezultato_matrica <- confusionMatrix(as.numeric(rezultatas$X1 > 0.5),
testavimas_y)
```

Matrica pateikia daug parametrų reikšmių, tačiau autoriaus manymu pakanka atsižvelgti į tris parametrus: tikslumą (angl. *accuracy*), jautrumą (angl. *sensitivity*), specifiškumą (angl. *specificity*). Šių parametrų reikšmės pateikiamos lentelėje žemiau.

1 lentelė. Apgaulingų transakcijų atpažinimo modelio verifikavimo rezultatai

Tikslumas	Jautrumas	Specifiškumas
0,9527	0,9801	0,9942

Gauti rezultatai reiškia, kad modelis dauguma atvejų teisingai nuspėja, ar transakcija apgaulinga. Vis dėlto specifiškumas yra didesnis už jautrumą, o tai reiškia, kad proporcingai rečiau apgaulingos transakcijos yra identifikuojamos kaip apgaulingos, negu tikros transakcijos laikomos tikromis. Kitaip tariant, siekiant efektyviau atpažinti apgaulingas transakcijas, reikia papildomai apmokyti modelį, arba naudoti kitus algoritmus, kad jautrumas būtų didesnis už specifiškumą. Tačiau tai yra siūlymas kitiems moksliniams tyrimams, šio darbo kontekste tolimesnė optimizacija nevykdoma.

2.5. Techninė analizė

Šiame poskyryje remiantis etaloninio bandymo karkasu „TPCx-BB“ sudaromos techninės duomenų gavybos paslaugų analizės gairės.

„TPCx-BB“ matuoja didžiųjų duomenų analitikos paslaugų efektyvumą, akcentuojant duomenų apimtį, spartą ir įvairovę. Šis karkasas apima apie 30 užklausų, padengiančių skirtingus analizės aspektus[WBC+17]. Šios užklausos nėra viešinamos, todėl norint bent iš dalies atkartoti matavimus yra poreikis sukurti alternatyvias užklausas. Taip pat šio karkaso dokumentacijoje apibrėžiamos metrikos: našumo metrika, kaina, sistemos prieinamumas ir energijos sąnaudos. Verta išsamiau paaiškinti šias metrikas, nes jomis remiantis bus atliekama techninė analizė:

1. Našumas. Ši metrika susideda iš smulkesnių kriterijų, reprezentuojančių apkrovą, galią, pralaidumą ir užklausų apdorojimo laiką. Parašytas skirtingo sudėtingumo 25 užklausos SQL kalba. Šiam žingsniui paimtas atskiras duomenų rinkinys, atviro kodo saugyklos „GitHub“ įrašai. Duomenų rinkinio struktūra - viena lentelė pavadinimu „bigquery-public-data.samples.github_timeline“. Esybių atributų yra daug, todėl bus išvardyti tik esminiai: „repository_url“, „repository_has_downloads“, „type“. Toliau pateikiamos keletas užklausų:

```
SELECT repository_url, repository_has_downloads, repository_owner
FROM 'bigquery-public-data.samples.github_timeline WHERE
repository_has_downloads IS TRUE DESC LIMIT 1000

SELECT * FROM 'bigquery-public-data.samples.github_timeline
WHERE repository_has_downloads IS TRUE
AND repository_owner LIKE '%react%' ASC LIMIT 1000
```

2. Kaina. Skaičiuojamas kainos ir našumo santykis. Šiai metrikai svarbus anksčiau minėtos našumo metrikos vertė ir paslaugos vartojimo kaina.
3. Sistemos prieinamumas. Apskaičiuojama, kiek kartų sistema nesugebėjo atlikti reikiamos procedūros ir ar yra specifinių laiko intervalų, kuomet pastebimas sistemos prieinamumo mažėjimas.

3. Palyginamoji analizė

Šiame skyriuje atliekama debesijos duomenų gavybos paslaugų **MS, Google, Amazon** palyginamoji analizė. Pirmame poskyryje trumpai aprašomos paslaugos, nagrinėjama jų dokumentacija. Antrame poskyryje remiantis 2.4 poskyryje metodu atliekama duomenų gavyba ir vertinamos programos galimybės iš vartotojo perspektyvos. Analizuojant paslaugos praktiškumą vertinama funkcionalumą ir paslaugos kokybę. Trečiame poskyryje atliekama našumo analizė. Ketvirtame poskyryje apibendrinami atliktų analizių rezultatai.

3.1. Dokumentacijos analizė

1. „Microsoft Azure HDInsight“ - debesų kompiuterijos paslauga, įgalinanti vartotojus naudotis populiariaisiais atviro kodo analitikos įrankiais „Apache“ - pavyzdžiui „Hadoop“. „Microsoft Azure HDInsight“ įgalina įgyvendinti daugelį tipinių scenarijų, susijusių su didžiaisiais duomenimis ir duomenų gavyba - ETL, DW, mašininį mokymąsi ir daiktų internetą. Ši paslauga pasirinkta dėl apmokymų prieinamumo, pozicijos rinkoje, gausios dokumentacijos bei galimybės nemokamai pasinaudoti paslauga[Mic].

2 lentelė. Microsoft Azure HDInsight ypatybės

Ypatybė	Paaiškinimas
Pritaikyta debesijai	Galima kurti optimizuotus klasterius platformoms „Apache Hadoop“, „Apache Spark“, „Apache Interactive Query“, „Apache Kafka“, „Apache Storm“, „Apache HBase“. „Microsoft Azure HDInsight“ pateikia paslaugų lygmens susitarimą (angl. *Service Level Agreement*) visoms vartotojo sukurtoms produkcijos aplinkoms.
Kaina ir plečiamumas	„Microsoft Azure HDInsight“ palaiko dinamišką resursų paskyrimą, priklausomai nuo poreikio kinta apkrovų ir susijusių resursų dydis. Norint minimizuoti kaštus, galima automatizuoti ir klasterių sukūrimą, kad duomenų apdorojimo procesas apimtų ir klasterio sukūrimą bei ištrinimą.
Saugumas ir atitiktis	„Azure Virtual Network“, šifravimas, bei integravimas į „Azure Active Directory“ apsaugo vartotojo intelektinę nuosavybę ir duomenis. „Microsoft Azure HDInsight“ atitinka populiariausius pramonės ir vyriausybės atitikties standartus.
Stebėjimas	„Microsoft Azure HDInsight“ integruojasi su „Azure Monitor logs“. Pateikiama viena sąsaja, leidžianti stebėti visus vartotojo klasterius.
Produktyvumas	„Microsoft Azure HDInsight“ leidžia naudoti integruotas programavimo aplinkas darbui su Spark ir Hadoop platformomis. Suderinama su tokiomis programomis/aplinkomis kaip „Visual Studio“, „VSCode“, „Eclipse“, „IntelliJ“, „Zeppelin“ ir „Jupyter“ ir palaiko .NET, Python, Java, Scala ir R programavimo kalbų naudojimą.
Išplečiamumas	Klasteriai gali būti išplėsti papildomais komponentais („Hue“, „Presto“ ir t. t.) naudojant skriptus, pridedant mazgus (angl. *edge nodes*) arba integruojant su kitais sertifikuotais didžiųjų duomenų sprendimais.

2. „Google Cloud Dataproc“ - debesų kompiuterijos paslauga, kuri suteikia galimybę valdyti

„Apache Hadoop“, „Apache Spark“ klasterius. Ji panaudoja kitas „Google Cloud Platform“ paslaugas, kad užtikrintų galingą ir visapusišką platformą duomenų apdorojimui, mašininiam mokymuisi bei analizei[Gooa]. „Google Cloud Dataproc“ aprūpina vartotoją valdymo ir integravimo mechanizmais, kurie koordinuoja klasterių gyvavimo ciklą, valdymą ir organizavimą. Ši paslauga pasirinkta dėl didelio mokomosios medžiagos kiekio, aiškios dokumentacijos bei galimybės nemokamai pasinaudoti paslauga. Lentelėje 3 surašytos „Google Cloud Dataproc“ ypatybės.

3 lentelė. Google Cloud Dataproc ypatybės

Ypatybė	Paaiškinimas
Klasterių valdymas	Klasterio diegimas, žurnalizavimas ir stebėjimas yra automatiškai valdomi. Klasterių dydis keičiamas priklausomai nuo reikalingo resursų kiekio.
Debesijos suderinamumas	Standartiškai paslauga integruota su tokiomis debesijos paslaugomis: Cloud Storage, BigQuery, Cloud Bigtable, Stackdriver Logging, Stackdriver Monitoring ir AI Hub.
Prieinamumas	Yra galimybė paleisti klasterius didelio prieinamumo režime. Naudojami keli pagrindiniai mazgai, o užduotys nustatomos perkrovai nesekmės atveju, kad klasteris ir užduotys būtų prieinamos visą laiką.
Verslo saugumas	Saugumui užtikrinti galima įjungti papildomą saugumo konfigūraciją kuriant klasterį - Hadoop Secure Mode. Taip pat „Google Cloud Dataproc“ apima standartinius visoms Google Cloud debesijos paslaugoms taikomus saugumo sprendimus, tokius kaip duomenų ramybėje šifravimas (angl. *at-rest encryption*), „OS Login“, „VPC Service Controls“ ir „Customer Managed Encryption Keys“.
Numatytas ištrynimas	Kad būtų išvengtos sąskaitos už nenaudojamą klasterį, siūloma naudoti numatytąjį ištrynimą, įgalinantį ištrinti klasterį - kai šis nenaudojamas nustatytą laiko intervalą, arba konkrečiu laiko momentu, arba praėjus nustatytam laiko periodui.
Klasterio konfigūravimas	Numatytas - automatinis, tačiau esant poreikiui galima keisti daugelį klasterio savybių rankiniu būdu
Papildomi komponentai	Yra galimybė integruoti papildomus komponentus, pateikiančius pilnai sukonfigūruotas aplinkas Zeppelin, Druid, Presto ir kitiems atviro kodo komponentams, skirtiems darbui su Hadoop ir Spark ekosistemomis.
Klasterio atvaizdas	Klasteriai gali būti aprūpinti jau paruoštu atvaizdu, kuriame įdiegti Linux operacinės sistemos paketai.
Komponentų	Component Gateway įgalina saugiai, vienu paspaudimu pasiekti „Google Cloud Dataproc“ komponentų, veikiančių sąsaja klasteryje, sąsajas.
Darbo eigos šablonai	Sąsaja „Workflow Templates“ įgalina lanksčiai ir paprastai kurti ir vykdyti darbo eigas (angl. *workflows*). Workflow Template tai pernaudojama darbo eigos konfigūracija, apibrėžianti užduočių eiliškumą, pavaizduojamą grafu.

3. „Amazon Elastic MapReduce“ - pritaikyta debesijai (angl. *cloud-native*) didžiųjų duomenų

platforma, naudojanti atviro kodo įrankius „Apache Spark“, „Apache Hive“, „Apache HBase“ ir kitus, kartu su „Amazon Elastic Cloud Compute“ ir „Amazon Simple Storage Service“ plečiamumu siūlo duomenų analitikams galingą variklį, sugebantį atlikti petabaitų rango duomenų analizę ženkliai pigiau negu standartinės (angl. *on-premises*) paslaugos ir klasteriai[Ama]. Lentelėje 4 surašytos Amazon Elastic MapReduce ypatybės.

4 lentelė. Amazon Elastic MapReduce ypatybės

Ypatybė	Paaiškinimas
Paprastumas	Paslauga pasirūpina reikiamais mazgais, infrastruktūros paruošimu, Hadoop konfigūravimu ir kitais klasterio nustatymais.
Kaina	Mokama už kiekvieną panaudotą sekundę su minimaliu vienos minutės mokesčiu.
Lankstumas	Yra galimybė aprūpinti klasterį tokiu mazgų skaičiumi, kokio reikia apdoroti duomenims - tiek rankiniu būdu, tiek automatiškai, naudojant paslaugą „Auto Scaling“ - kuri nustato klasterio dydį pagal apkrovą. Skaičiavimas ir duomenų saugojimas yra atskiriami, todėl kiekviena mazgų grupė gali būti plečiama nepriklausomai viena nuo kitos.
Patikimumas	Paslauga pritaikyta debesijai, nuolat stebimas klasterio veikimas - nesėkmingos užduotys vykdomos iš naujo, o nepakankamai greitai veikiantys mazgai pakeičiami kitais.
Saugumas	„Amazon Elastic MapReduce“ automatiškai sukonfigūruoja „Amazon Elastic Cloud Compute“ ugniasienės nustatymus, susijusius su prieiga prie mazgų tinklu ir paleidžia klasterius Amazon Virtual Private Cloud, logiškai izoliuotame vartotojo apibrėžtame tinkle.
Lankstumas	Yra prieiga prie kiekvieno dirbančio mazgo, kiekviename jame galima įdiegti papildomus įskiepius ir komponentus, taip pat galima pakoreguoti klasterio įkrovą. Naudojant Amazon Machine Images - sistemos atvaizdus - paleisti klasterius su nestandartiniais operacinės sistemos nustatymais ir keisti klasterių konfigūraciją neperkraunant jo.

3.2. Duomenų gavybos procesas

Šiame poskyryje realizuojamas duomenų gavybos procesas, kurio tikslas yra sukurti modelį, nustatantį apgaulingas transakcijas bei šio modelio validacija. Pirminis programos tekstas su paaiškinimais pateiktas 2.4 poskyryje. Žemiau išvardijami duomenų gavybos proceso įgyvendinimai naudojant „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“:

1. „Microsoft Azure HDInsight“. Pirmą kartą dirbant su „Microsoft Azure HDInsight“ gali kilti neaiškumų dėl gausaus nustatymų skaičiaus. Vis dėlto remiantis mokymo literatūra pavyzdinio projekto kūrimo žingsniai padeda paruošti aplinką. Aplinkos paruošimo metu reikia sukurti klasterį, resursų grupę, duomenų saugyklos paskyrą.

   Kuriant klasterį reikalaujamos dvi prisijungimo paskyros - antroji skirta prisijungti prie klasterio naudojant saugųjį tinklo protokolą SSH. Jungiantis prie klasterio valdymo konsolės buvo reikalaujama naudoti antrąją, kas suponuoja papildomą resursų apsaugą.

   Pastebėtas trukūmas iš vartotojo patogumo perspektyvos - ruošiant aplinką, jei sukuriamos

duomenų saugyklos pavadinimas nevalidus arba jau užimtas, apie tai pranešama tik pradėjus klasterio diegimo (angl. *deployment*) procesą, todėl sugaištama apie minutę. Klasterio sukūrimo laikas apie 20 minučių.

Visi užduoties aplinkai sukurti resursai (šiuo atveju klasteris) pateikiami „valdymo skydelyje“ (angl. *dashboard*) ir vartotojas aiškiai mato kiek mazgų sudaro klasterį, bei kiek branduolių bei mazgų dedikuota konkrečiai grupei.

Užduoties atlikimui būtina sąlyga, kurią turi užtikrinti paslauga yra galimybė vykdyti kodą, kuris atliks užduotis, susijusias su duomenų gavyba. Su šia paslauga prieinama naršyklėje integruota kūrimo aplinka „RStudio community edition for ML Services“. Tai išlaisvina vartotoją nuo papildomų procedūrų, susijusių su programavimo aplinkos diegimu ir konfigūravimu.

Modelio validacijos reikšmės pateiktos lentelėje žemiau.

5 lentelė. Apgaulingų transakcijų aptikimo modelio validacijos rezultatai naudojant „Microsoft Azure HDInsight“

Kriterijus	Tikslumas	Jautrumas	Specifiškumas
Reikšmė	0,9430	0,9617	0,9957

2. „Google Cloud Dataproc“. Kuriant projektą „Google Cloud Dataproc“ valdymo skydelyje, reikia išreikšinai įjungti atsiskaitymą šiam projektui. Klasterio kūrimas pagal rekomendacijas trunka apie 6 minutes. Parinkta konfigūracija - 1 vedantysis (2 virtualūs procesoriai, 7.5 gigabaitų operatyviosios atminties) ir 5 vedamieji mazgai (kiekvienas turi po 1 virtualų procesorių ir 3.75 gigabaitų operatyviosios atminties).

   R programavimo aplinkos paruošimas nėra automatizuotas. Reikia iš savo kompiuterio nuotoliniu būdu, naudojant „Gcloud“ ir SSH protokolą, prisijungti prie klasterio vedančiojo mazgo. Prisijungus, reikia suinstaliuoti papildomus paketus Linux operacinei sistemai. Galiausiai, parsisiųsti RStudio Server. Sukuriamas operacinės sistemos vartotojas, kuris galės prisijungti prie RStudio Server programos. Taip pat reikia sukonfigūruoti naršyklę, kad galima būtų prisijungti prie RStudio aplinkos naudojant SSH tunelį su SSH SOCKS tarpiniu serveriu.

   Modelio validacijos reikšmės pateiktos lentelėje žemiau.

6 lentelė. Apgaulingų transakcijų aptikimo modelio validacijos rezultatai naudojant „Google Cloud Dataproc“

Kriterijus	Tikslumas	Jautrumas	Specifiškumas
Reikšmė	0,9683	0,9811	0,9925

3. „Amazon Elastic MapReduce“. „Amazon Elastic MapReduce“ suteikia galimybę naudotis R programavimo aplinka viename iš darbinių mazgų. Sukurti klasterį galima per „Amazon Elastic MapReduce“ konsolinę sąsają. Privalu pasirinkti programinės įrangos versiją, ir klasterio tipą - pavyzdžiui Spark. Empirinio tyrimo metu visi kiti nustatymai palikti numatytieji. Klasteris paleistas per AWS CLI, alternatyviai galima paleisti Java aplinkoje.

Turint veikiantį klasterį, naudojant AWS CloudFormation šabloną, sukuriamas atskiras darbinis mazgas, kuriame veiks RStudio ir užkraunami duomenys į duomenų bazę - šiuo atveju failas su duomenimis įkeliamas iš lokalaus kompiuterio. Atidaromas RStudio per CloudFormation meniu.
Modelio validacijos reikšmės pateiktos lentelėje žemiau.

7 lentelė. Apgaulingų transakcijų aptikimo modelio validacijos rezultatai naudojant „Amazon Elastic MapReduce“

Kriterijus	Tikslumas	Jautrumas	Specifiškumas
Reikšmė	0,9516	0,9749	0,9914

3.3. Našumo analizė

Šiame skyriuje atliekama duomenų gavybos paslaugų „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“ našumo analizė. Kiekvienoje paslaugoje pamatuojama, kaip greitai įvykdomos užklausos. Visų užklausų apdorojimo laikai agreguojami į 5 reikšmes: vidurkį, pirmą, antrą ir trečią kvartilius bei 90-ąjį kvantilį. Žemiau aprašoma užklausų apdorojimo vykdymo laikų analizė išvardytoms paslaugoms:

1. „Microsoft Azure HDInsight“. Įvykdžius visas užklausas gauti agreguoti rezultatai pateikiami lentelėje žemiau:

8 lentelė. „Microsoft Azure HDInsight“ našumo rezultatai

Kriterijus	Vidurkis	1-as kvartilis	2-as kvartilis	3-as kvartilis	90-asis kvantilis
Reikšmė	3,17256	2,455	2,91	3,975	5,606

2. „Google Cloud Dataproc“. Įvykdžius visas užklausas gauti agreguoti rezultatai pateikiami lentelėje žemiau:

9 lentelė. „Google Cloud Dataproc“ našumo rezultatai

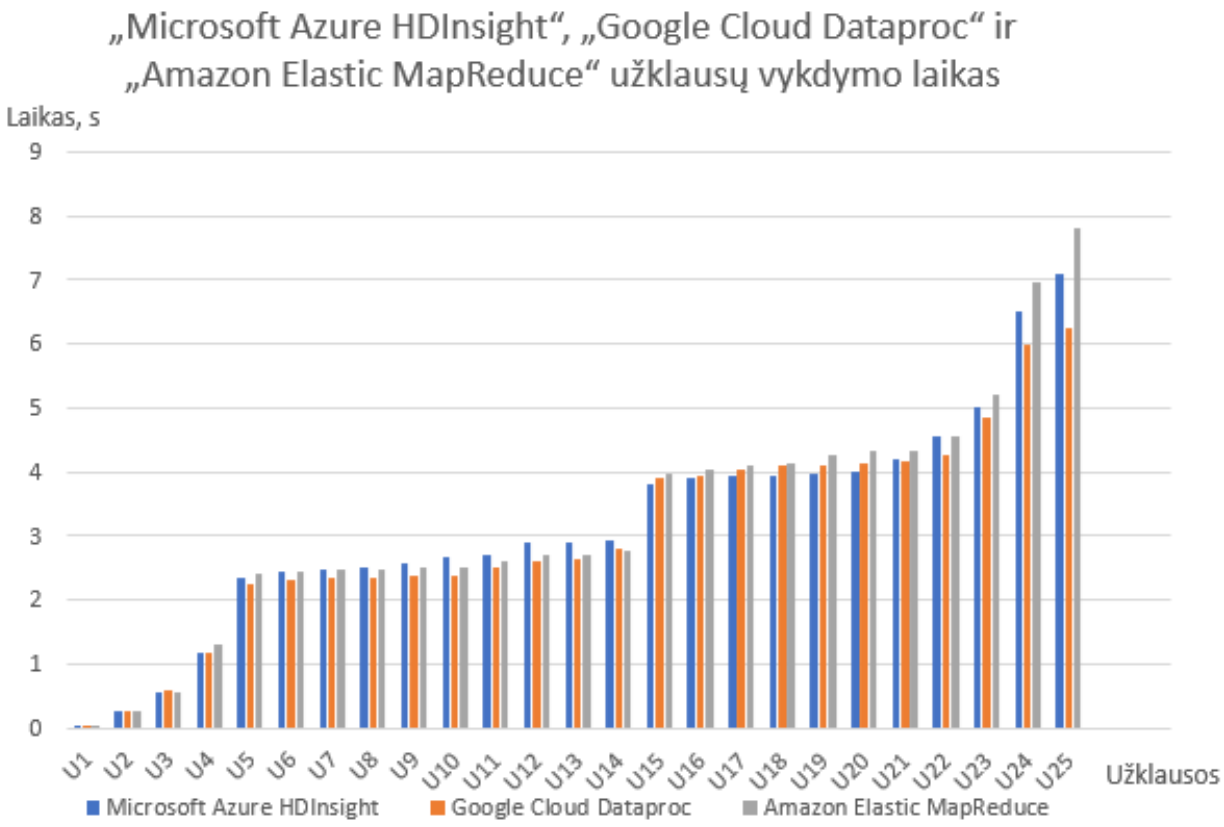
Kriterijus	Vidurkis	1-as kvartilis	2-as kvartilis	3-as kvartilis	90-asis kvantilis
Reikšmė	3,055	2,325	2,63	4,115	5,3

3. „Amazon Elastic MapReduce“. Įvykdžius visas užklausas gauti agreguoti rezultatai pateikiami lentelėje žemiau:

10 lentelė. „Amazon Elastic MapReduce“ našumo rezultatai

Kriterijus	Vidurkis	1-as kvartilis	2-as kvartilis	3-as kvartilis	90-asis kvantilis
Reikšmė	3,25496	2,436	2,71	4,285	5,906

Visi užklausų apdorojimo greičiai, priklausomai nuo duomenų gavybos paslaugos, pateikti grafike žemiau:



3 pav. Užklausų apdorojimo greičiai

Grafike matoma, kad beveik visais atvejais greičiausiai užklausa apdorojama „Google Cloud Dataproc“ paslaugoje. Kita vertus, „Amazon Elastic MapReduce“ beveik visais atvejais lėčiausiai apdorojo užklausas. Todėl užklausų apdorojimo kontekste geriausias įvertinimas tenka „Google Cloud Dataproc“, blogiausias - „Amazon Elastic MapReduce“.

3.4. Apibendrinimas

Remiantis ankstesnių poskyrių rezultatais ir apibrėžtais duomenų gavybos paslaugų palyginimo aspektais sudaroma palyginimo lentelė. Lentelėje žemiau palyginamos duomenų gavybos paslaugos „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“.

11 lentelė. „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“ palyginimas pagal aspektus

Aspektas	„Microsoft Azure HDInsight“	„Google Cloud Dataproc“	„Amazon Elastic MapReduce“
Funkcionalumas	Duomenų importavimo galimybė naudojant „HDFS“, „Azure Data Lake Storage“. Duomenys gali būti eksportuojami naudojant „Curl“ įrankį. Duomenų gavybos kodo modifikavimui prieinama aplinka „RStudio“. Duomenų transformavimui ir paruošimui naudojama „Azure Data Factory“.	Duomenų importavimo galimybės naudojant „Scoop“ ir „HDFS“. Duomenys eksportuojami per „Google Cloud Storage“. Kodo modifikavimą įgalina „RStudio“. Duomenų transformacijos operacijos atliekamos su „Cloud Dataflow“.	Duomenų importavimo galimybės naudojant „Amazon S3“, „AWS Direct“ Connect“. Duomenys eksportuojami naudojant „Amazon S3“, „DynamoDB“. Kodas modifikuojamas aplinkoje „RStudio“. Duomenų transformacija vykdoma naudojant „Amazon EC2“.
Vartotojo sąsaja	Išsamus vartotojo gidas padeda surasti reikiamas funkcijas, nors paslaugos valdymo skydelyje išdėstymas nėra itin intuityvus. Nėra vizualizuoto programavimo galimybės. Atliekant klasterio diegimo procesą, vediklis padeda atlikti žingsnius.	Intuityvi vartojo sąsaja. Instrukcijos ir vedlys įgalina greitai išmokti naudotis paslauga, tiek diegiant klasterį, tiek paleidžiant tekstinę aplinką nekyla sunkumų.	Nėra aiškių instrukcijų kaip pasiekti reikiamas funkcijas, pvz. „RStudio“ aplinką. Net atlikus duomenų gavybos procesą nepagerėjo vartotojo patirtis.
Integralumas	Suderinama su atviro kodo komponentais „Hadoop“, „Spark“, „Hive“, „LLAP“, „Kafka“, „Storm“, „HBase“ bei programavimo kalbomis „Java“, „Python“, „.NET“ „Go“, „R“ ir „Scala“.	Galima panaudoti komponentus „Zeppelin“, „Druid“, „Presto“, „RStudio“, „Hadoop“ ir „Spark“.	Yra galimybė naudoti „RStudio“, „Presto“, „Druid“.

Rezultatai ir išvados

Išvardijami darbo rezultatai. Išanalizuotas duomenų gavybos konceptas ir nustatyti duomenų gavybos proceso sudedamieji žingsniai. Apibrėžti debesų kompiuterijos paslaugų, skirtų duomenų gavybai, palyginimo aspektai - funkcionalumas, vartotojo sąsaja, integralumas ir našumas. Apgaulingų transakcijų, atliktų kreditinėmis kortelėmis, aptikimo modelis realizuotas ir validuotas panaudojant duomenų gavybos paslaugas „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“. Išnagrinėjus išvardytų paslaugų dokumentaciją, atsižvelgus į duomenų gavybos proceso įgyvendinimo metu gautą patirtį ir apibrėžtus palyginimo aspektus, atlikta duomenų gavybos paslaugų palyginamoji analizė. Praktiniu pavyzdžiu parodyta duomenų gavybos ir ją atlikti įgalinančių paslaugų nauda realiame pasaulyje - aptikti ir blokuoti apgaulingas transakcijas bankų sistemose.

Išvados:

1. „Microsoft Azure HDInsight“, „Google Cloud Dataproc“ ir „Amazon Elastic MapReduce“ įgalina programiškai atlikti duomenų gavybos procesą, tačiau skiriasi tam reikalingos programavimo aplinkos paruošimo vartotojo sąsajos kokybė. Geriausiai šiuo aspektu vertinama „Microsoft Azure HDInsight“ paslauga, o blogiausiai - „Amazon Elastic MapReduce“.
2. Našumo analizės rezultate nustatyta, kad greičiausiai užklausos apdorojamos naudojant „Google Cloud Dataproc“ paslaugą, lėčiausiai - „Amazon Elastic MapReduce“.
3. Įvertinus visus aspektus - funkcionalumą, vartotojo sąsają, našumą ir kainą - optimalus pasirinkimas iš išvardytų paslaugų yra „Google Cloud Dataproc“.
4. Apgaulingų transakcijų aptikimo modelio validacijos metu nustatyta, kad jautrumas mažesnis už specifiškumą, o siekiamybė yra atvirkščia situacija, todėl rekomenduojama keisti duomenų gavybos proceso metodų „RandomForest“ ir „Synthetic Minority Over-sampling TEchnique“ parametrus arba naudoti kitus metodus.

Literatūra

[AFG+09] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph ir k.t. Above the Clouds: A Berkeley View of Cloud Computing. *Technical Report No. UCB/EECS-2009-28*:1, 2009.

[Ama] Amazon. Amazon Elastic MapReduce. `https://aws.amazon.com/emr/`. Tikrinta 2021-05-04.

[CBH+02] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall ir W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*:321–357, 2002.

[CCS12] Hsinchun Chen, Roger H. L. Chiang ir Veda C. Storey. BUSINESS INTELLIGENCE AND ANALYTICS : FROM BIG DATA TO BIG IMPACT. *MIS Quarterly*, 36:1666, 2012.

[Cen] Barcelona Supercomputing Centre. ALOJA Big Data Benchmarking platform. `https://github.com/Aloja/aloja`. Tikrinta 2021-05-01.

[DSS+17] *A Comparison of Contemporary Data Mining Tools*. XVII International Scientific Conference on Industrial Systems, 2017.

[FPS96a] Usama Fayyad, Gregory Piatetsky-Shapiro ir Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17:38, 1996.

[FPS96b] Usama Fayyad, Gregory Piatetsky-Shapiro ir Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17:41, 1996.

[Gooa] Google. Google Cloud Dataproc. `https://cloud.google.com/dataproc/`. Tikrinta 2021-05-04.

[Goob] Google. Google Duomenų skaidrumas. `https://safety.google/privacy/data/`. Tikrinta 2021-05-04.

[GP03] C. Giraud-Carrier ir O. Povel. Characterising Data Mining software. *Intelligent Data Analysis*, 7:181–192, 2003.

[Jos13] Madhuri V. Joseph. Data Mining and Business Intelligence Applications in Telecommunication Industry. *International Journal of Engineering and Advanced Technology (IJEAT)*, 2, 2013.

[KEG+98] M.A. King, John Fletcher Elder, B. Gomolka, and E. Schmidt. Evaluation of fourteen desktop data mining tools. 3, 1998.

[KT] Hian Chye Koh ir Gerald Tan. Data Mining Applications in Healthcare. *Journal of Healthcare Information Management*, 19:64–72.

[KU18] Kaggle ir Machine Learning Group ULB. Credit Card Fraud Detection, 2018.

[Lan01a] Doug Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *APPLICATION DELIVERY STRATEGIES*, 2001.

[Lan01b] Doug Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *APPLICATION DELIVERY STRATEGIES*:1–2, 2001.

[Lan01c] Doug Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *APPLICATION DELIVERY STRATEGIES*:2, 2001.

[LS01] Sang Jun Lee ir Keng Siau. A review of data mining techniques. *Industrial Management & Data Systems*:41–46, 2001.

[MA12] Kazi Imran Moin ir Qazi Baseer Ahmed. Use of Data Mining in Banking. *International Journal of Engineering Research and Applications (IJERA)*, 2:738–742, 2012.

[MB12] Andrew McAfee ir Erik Brynjolfsson. Big Data: The Management Revolution. *Harvard Business Review*:5, 2012.

[MB14] A. McAfee ir E. Brynjolfsson. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35:138, 2014.

[Mic] Microsoft. Microsoft Azure HDInsight. `https://azure.microsoft.com/en-in/services/hdinsight/`. Tikrinta 2021-05-04.

[MUN15] Mihaela MUNTEAN. Considerations Regarding Business Intelligence in Cloud Context. *Informatica Economică*, 19:55–67, 2015.

[NSC+16] P. Neves, B. Schmerl, J. Cámara ir J. Bernardino. Big Data in Cloud Computing: Features and Issues. *Proceedings of the International Conference on Internet of Things and Big Data*, 2016.

[PCJ+15] *Calibrating Probability with Undersampling for Unbalanced Classification*. 2015 IEEE Symposium Series on Computational Intelligence, 2015. IEEE.

[PMC17] Nicolas Poggi, Alejandro Montero ir David Carrera. Characterizing TCPBx-BB queries, Hive and Spark in multi-cloud. *Lecture Notes in Computer Science*, 10661, 2017.

[RGR18] David Reinsel, John Gantz ir John Rydning. The Digitization of the World From Edge to Core. *An IDC White Paper – #US44413318*:6, 2018.

[Sol13] Harshvardhan Solanki. Comparative Study of Data Mining Tools and Analysis with Unified Data Mining Theory. *International Journal of Computer Applications*:23–28, 2013.

[TPC] TPC. TPC-DS Specification V 1.0.0L. `http://www.tpc.org/tpc_documents_current_versions/pdf/tpc-ds_v2.1.0.pdf`. Tikrinta 2021-05-02.

[TPC21] TPC. TPC Express Big Bench Standard Specification V 1.5.0. `http://tpc.org/tpcx-bb/default5.asp`, 2021. Tikrinta 2021-05-02.

[VC15] Liwen Vaughan ir Yue Chen. Data Mining From Web Search Queries: A Comparison of Google Trends and Baidu Index. *JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY*, 66:13–22, 2015.

[VD14] Vikas Verma ir Sunil Dhawan. Methodology for Selection of a Data Mining Tool. *International Journal of Software & Hardware Research in Engineering*, 2:189–192, 2014.

[VH00] Howard Veregin ir Gary Hunter. Data Quality Measurement and Assessment. *NCGIA Core Curriculum in Geographic Information Science*, 2000.

[WAA+] Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi ir Emad M. Al-Shawakfa. A Comparison Study between Data Mining Tools over some Classification Methods. *(IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*:18–22.

[WBC+17] Kebing Wang, Bianny Bian, Paul Cao ir Mike Riess. Lessons learned using the TPCx-BB Express Benchmark for end-to-end big data clusters, 2017.

Sąvokų apibrėžimai

Atsitiktinių medžių rinkinys - klasifikacijos metodas, naudojantis sprendimų medžius.

Koncepto hierarchija - atvaizdžių iš žemesnio lygių atributų rinkinio į aukštesnio lygio atributų rinkinį seka.

Kubinė duomenų agregacija - duomenų redukcija, keičiant duomenų formatą į daugiamatį masyvą.

Paslaugų lygmens susitarimas - susitarimas tarp vartotojo ir paslaugų teikėjo, apibrėžiantis paslaugų rūšį ir numatantis tų paslaugų teikimo bei jų kokybės garantijas, taip pat vartotojo eksploatacinius bei finansinius įsipareigojimus.

Pažingsninio parinkimo algoritmas - algoritmas skirtas automatiškai atrinkti geriausius arba blogiausius atributus.

Principinė komponentų analizė - duomenų dimensijų mažinimo technika, paremta kovariacijos matrica.

Santrumpos

Amazon EC2 - (angl. *Amazon Elastic Cloud Compute*) debesų kompiuterijos kompiuterinio skaičiavimo paslauga.

CSV - (angl. *Comma-separated Values*) kableliais atskirtos reikšmės, duomenų saugojimo formatas.

ETL - (angl. *Extract, Transform, Load*) duomenų gavybos procesas, leidžiantis surinkti duomenis iš skirtingų šaltinių ir konsoliduoti į centralizuotą saugojimo vietą.

HDFS - (angl. *Hadoop Distributed File System*) - paskirstyta failų sistema.

OLAP - (angl. *OnLine Analytical Processing*) tai technologija įgalinanti kurti verslo analitikos taikomąsias programas. Naudojama duomenų gavybai, interaktyviai ataskaitų peržiūrai, kompleksiniams analitiniams skaičiavimams.

SMOTE - (angl. *Synthetic Minority Oversampling TEchnique*) dirbtinė pavyzdžių sukūrimo technika.

SSH - (angl. *Secure SHell*) protokolas, tinkle užtikrinantis saugų nuotolinį prisijungimą ir kitas saugaus tinklo paslaugas.

TPCx-BB - (angl. *Transaction Processing Performance Council Express - Big Bench*) didžiųjų duomenų analitinių įrankių etaloninio bandymo sistema.