VILNIAUS UNIVERSITETAS

EKONOMIKOS IR VERSLO ADMINISTRAVIMO FAKULTETAS

EKONOMINĖS INFORMATIKOS KATEDRA

Marius STULGINSKAS

Verslo informacinių sistemų programa

DUOMENŲ TINKLO ARCHITEKTŪROS POREIKĮ LEMIANTYS VEIKSNIAI ORGANIZACIJOSE

BAKALAURO DARBAS

Leidžiama ginti _____

(parašas)

Katedros vedėjas **prof. dr. R. Skyrius**

Bakalaurantas _____

(parašas)

Darbo vadovas _____

(parašas)

Dr. Asistentas Michail Kazimianec

Darbo įteikimo data: _____

Registracijos Nr. _____

Vilnius, 2022

Turinys

LENTELIŲ SĄRAŠAS

SANTRUMPŲ SĄRAŠAS

ETL *(Extract Transfrom Load)* – išgavimas, transformavimas, užkrovimas

ELT *(Extract Load Transform)* – išgavimas, užkrovimas, transformavimas

SQL *(Structured Query Language)* – struktūruota užklausų kalba

OLAP *(Online-Analytical Processing)* – internetinis-analitinis duomenų apdorojimas

OLTP *(Online Transactions Processing)* – internetinis transakcijų apdorojimas

CPU *(Central Processing Unit)* – centrinis procesorius

RAM *(Random-Access Memory)* – operatyvioji atmintis

MPP *(Massively Parallel Processing)* – masyvus paralelinis apdorojimas

SSH *(Shell Secure)* – tinklo protokolas

CRM *(Customer Relationship Management)* – santykių su klientais valdymo sistema

ERP *(Enterprise Resource Planning)* – verslo valdymo sistema

API *(Application Programming Interface)* – aplikacijų programavimo sąsaja

KPI *(Key Performance Indicator)* – pagrindinis veiklos rodiklis

TDWI *(Transforming Data With Intelligence)* – duomenų transformavimas paremtas įžvalgomis

UML *(Unified Modeling Language)* – vieninga modeliavimo kalba

TLS *(Transport Layer Security)* – kriptografinis protokolas

JSON *(Java Script Object Notation)* – atviro standarto formatas

Įvadas

Šiais laikais praktiškai kiekvienam yra suprantama, kad duomenys yra turtas. Duomenų kiekiui sparčiai augant atsiranda didžiųjų duomenų iššūkis. Tokiam duomenų kiekio valdymui paprastų saugojimo technologijų nebepakanka, todėl atsiranda duomenų ir informacijos valdymo architektūros. Duomenų ir informacijos valdymas yra vienas kitą papildantys komponentai, o su kokybiška duomenų valdymo strategija, padidėja tikimybė pasiekti užsibrėžtų tikslų. Tačiau jau nuo seno žmonės suprato, kad norint duomenis naudoti, visų pirma reikia turėti kur juos saugoti ir laikyti. Taip atsirado duomenų bazės idėja. Duomenų bazės ištakos siekia bibliotekas, vyriausybinius, verslo ir medicininius įrašus dar prieš išrandant kompiuterius. Kai žmonės suprato, kad reikia turėti priemonių duomenims saugoti, jie bandė rasti būdų, kaip saugoti, indeksuoti ir gauti duomenis. Atsiradus kompiuteriams, duomenų bazių pasaulis sparčiai pasikeitė, todėl duomenų bazėse esančius duomenis kaupti ir tvarkyti tapo paprasta, ekonomiška ir mažiau vietos užimančia užduotimi. (Seymour ir Berg, 2012). Laikui bėgant, o duomenų kiekiui vis sparčiau augant, tobulėjo ir technologijos, skirtos duomenims apdoroti, saugoti ir analizuoti. Duomenų saugykla ir duomenų ežeras gana nemažą laiko tarpą buvo didelė jėga duomenų valdymo pasaulyje, tačiau didėjantis duomenų kiekis ir kompleksiškumas sukelia vis daugiau iššūkių juos valdant. Tiek duomenų saugykla, tiek duomenų ežeras, turi centralizuotą duomenų saugojimo repozitoriją, o tai sukelia tokius iššūkius kaip: sunkus duomenų savininkų atsekamumas (keičiasi duomenų valdytojai); sunkus duomenų šaltinių atsekamumas (didelis duomenų kiekis, skirtingi duomenų rinkiniai iš skirtingų vietų); visi duomenys gali būti vieno žmogaus ar atskiro departamento rankose, kas lemia, kad duomenys gali būti suprasti ne taip, niekas gali neprisiimti atsakomybės už duomenis arba neturėti pakankamai žinių tokius duomenis tinkamai interpretuoti; duomenų nuosavybės (*angl. Data Ownership*) fragmentiškumas (gali būti, kad tik dalis duomenų turi savo savininką). Tačiau būtent šioje vietoje, įžengia duomenų tinklo architektūros idėja (*angl. Data Mesh*). Ši architektūra apibrėžia decentralizuotą duomenų kaupimą ir valdymą, o tai padeda bent jau iš dalies išspręsti visas anksčiau išvardintas problemas, kurios kyla naudojant centralizuotą repozitoriją. Plačiau apie duomenų valdymo iššūkius, jų sprendimo būdus, skirtingų architektūrų skirtumus ir naudą bei veiksnius, kurie lemia duomenų tinklo architektūros poreikį, bus rašoma toliau šiame darbe.

Darbo tikslas - nustatyti esminius veiksnius, lemiančius duomenų valdymo problemas ir iššūkius.

Darbo uždaviniai:

1. Remiantis literatūros analize palyginti informacijos ir duomenų valdymo architektūras
2. Išnagrinėti klasikinę verslo įžvalgos ir duomenų valdymo architektūrą
3. Išnagrinėti kokią naudą organizacijai teikia efektyvus metaduomenų valdymas
4. Išnagrinėti su kokiomis pagrindinėmis problemomis susiduria organizacijos siekdamos efektyviai valdyti duomenis
5. Remiantis apklausos rezultatais, įvertinti įmonių pasirengimą ir poreikį pritaikyti duomenų tinklo architektūrą
6. Pateikti išvadas ir pasiūlymus/rekomendacijas

1. Duomenų ir informacijos valdymas

Duomenų valdymas yra praktiškai kertinis veiksnys, kuris nulemia organizacijos sėkmę. Dauguma puikiai supranta, kad kokybiškas ir efektyvus duomenų valdymas gali atnešti puikius rezultatus, tačiau toli gražu ne visi suvokia, kaip tinkamai tuos duomenis reikia valdyti. Organizacijos, turinčios patikimų ir kokybiškų duomenų apie savo klientus, produktus, paslaugas ir veiklą, gali priimti geresnius sprendimus nei tos, kurios neturi duomenų arba turi nepatikimus duomenis. (Data Management Body...)

Pats duomenų valdymas apimą daug procesų ir veiksnių, tačiau, kad procesai vyktų turi būti nustatyta tvarka ir taisyklės, kurių reikia laikytis, norint tinkamai vykdyti duomenų valdymo įgyvendinimą. Gana dažnai už tokią tvarką būna atsakinga kita duomenų valdymui gimininga šaka – duomenų valdymo taisyklių įgalinimas (*angl. Data Governance*). Šie du komponentai vienas su kitu yra glaudžiai susiję, tačiau duomenų valdymo taisyklių įgalinimas yra gyvybiškai svarbus duomenų valdymui. Paprastai tariant, duomenų valdymo taisyklių įgalinimas sukuria įvairias taisykles ir procedūras aplink duomenis, o duomenų valdymas yra atsakingas už tų taisyklių ir procedūrų įgyvendinimą ir pritaikymą duomenims. Įprastai, kad susidaryti aiškesnį vaizdą apie tai, kas yra duomenų valdymo taisyklių įgalinimas, tikslinga į tai pažvelgti per klausimų prizmę, t.y. į kokius klausimus duomenų valdymo taisyklių įgalinimas atsako. Keli pavyzdžiai būtų: „Kas turi nuosavybęs teisę į duomenis?“, „Kas ir kokius duomenis gali pasiekti?“, „Kokie saugumo parametrai yra nustatyti, siekiant apsaugoti duomenis ir jų privatumą?“, „Kokie duomenų šaltiniai leistini naudoti?“ ir panašiai. (Data Management vs...). Toliau šiame skyriuje bus apžvelgti duomenų valdymo veiksniai, kuriuos užtikrinus, organizacijos duomenų valdymas tamptų efektyvesnis ir kokybiškesnis.

Duomenų valdymas ir kokybė. Duomenų valdymas turi bendrų bruožų su kitomis turto valdymo formomis. Jis apima žinojimą, kokius duomenis organizacija turi ir ką su jais galima pasiekti, tada nustatant, kaip geriausiai panaudoti duomenų turtą organizacijos tikslams pasiekti. Kaip ir kiti valdymo procesai, duomenų valdymas turi suderinti strateginius ir veiklos poreikius. Šią pusiausvyrą geriausiai galima pasiekti laikantis tam tikrų principų, kuriais pripažįstamos svarbiausios duomenų valdymo ypatybės ir kuriais vadovaujamasi duomenų valdymo praktikoje. Galime teigti, kad duomenys yra unikalių savybių turintis turtas. Duomenys yra turtas, tačiau nuo kito turto jie skiriasi svarbiais aspektais, kurie daro įtaką jų valdymui. Labiausiai akivaizdi iš šių savybių yra ta, kad duomenys nėra sunaudojami, kai jie naudojami, kaip kad yra su finansiniu ir fiziniu turtu. Todėl jei duomenys vadinami turtu, tai reiškia, kad jie turi vertę. Nors yra duomenų kokybinės ir kiekybinės vertės matavimo metodų, dar nėra

tam skirtų standartų. Organizacijos, norinčios priimti geresnius sprendimus dėl savo duomenų, turėtų sukurti nuoseklius kiekybinio vertės nustatymo būdus. Jos taip pat turėtų įvertinti ir žemos kokybės duomenų sąnaudas, ir aukštos kokybės duomenų naudą. O, kad užtikrinti duomenų kokybės valdymą, organizacija privalo efektyviai valdyti visus savo duomenis. (Data Management Body...) Tačiau realaus gyvenimo duomenys dažnai būna „nešvarūs“: nenuoseklūs, besidubliuojantys, netikslūs, neišsamūs arba pasenę. Dėl netinkamų duomenų, nuolat gaunami klaidinantys arba neobjektyvūs analizės rezultatai ir sprendimai, prarandamos pajamos, patikimumas ir klientai. Dėl to atsiranda duomenų kokybės valdymo poreikis, kad būtų pagerinta duomenų kokybė, aiškiai apibrėžta jų paskirtis ir taip padidinta verslo procesų vertė. (Fan ir Geerts, 2012) Duomenų paskirties atitikimas yra pagrindinis duomenų valdymo tikslas, norėdamos valdyti kokybę, organizacijos turi užtikrinti, kad suinteresuotųjų šalių reikalavimai kokybei būtų tinkamai suprasti ir vykdomi. (Data Management Body...)

Metaduomenų svarba. Duomenų valdymas taip pat labai priklauso nuo metaduomenų valdymo. Kadangi duomenų negalima laikyti ar paliesti, norint suprasti, kas jie yra ir kaip juos naudoti, reikia apibrėžimo ir žinių metaduomenų forma. Metaduomenys gaunami vykdant įvairius su duomenų kūrimu, apdorojimu ir naudojimu susijusius procesus, įskaitant architektūrą, modeliavimą, tvarkymą, valdymą, duomenų kokybės valdymą, sistemų kūrimą, IT ir verslo operacijas bei analizę. (Data Management Body...) Laikui bėgant, vis daugiau organizacijų į metaduomenis pradeda žiūrėti ne tik kaip į „duomenis apie duomenis“, tačiau juos traktuoja kaip galimas jungtis tarp skirtingų duomenų šaltinių, duomenų savininkų ir jau egzistuojančių duomenų. Išsprendus duomenų atsekamumo problemą, duomenų valdymas ateityje turėtų tapti daug paprastesnis ir efektyvesnis. (Riley, 2017)

Planavimas. Norint valdyti duomenis, reikia planuoti. Duomenys kuriami daugelyje vietų ir perkeliami iš vienos vietos į kitą. Norint koordinuoti darbą ir išlaikyti galutinius rezultatus suderintus, reikia planuoti architektūrą ir procesus. (Data Management Body...)

Įgūdžiai ir kompetencijos. Duomenų valdymas yra daugiafunkcinis. Jam reikalingi įvairūs įgūdžiai ir kompetencijos. Viena komanda negali valdyti visų organizacijos duomenų, todėl duomenų valdymui reikalingi tiek techniniai, tiek netechniniai įgūdžiai ir gebėjimas bendradarbiauti. (Data Management Body...) Galime išskirti keletą kompetencijų, kurios įvardijamos, norint efektyviai valdyti duomenis: suprasti ir gebėti pritaikyti skirtingus duomenų kaupimo metodus; suprasti metaduomenų pritaikymo svarbą ir mokėti tinkamai juos panaudoti; mokėti tinkamai saugoti duomenis, kad išvengti duomenų nesutapimų; mokėti išsaugoti duomenų autentiškumą; vadovautis saugos reikalavimais; gebėti atsekti duomenų šaltinius ir duomenų savininkus ir panašiai. (Grillenberger ir Romeike, 2014)

Įmonės perspektyva. Duomenų valdymui reikalinga įmonės perspektyva. Kad duomenų valdymas būtų kuo veiksmingesnis, jis turi būti taikomas visoje įmonėje. (Data Management Body...)

Duomenų kintamumas. Duomenų valdymas turi atsižvelgti į įvairias perspektyvas. Duomenys yra kintantys, todėl duomenų valdymas turi nuolat tobulėti, kad neatsiliktų nuo duomenų kūrimo ir naudojimo būdų bei juos naudojančių duomenų vartotojų. (Data Management Body...)

Rizikų valdymas. Duomenų valdymas apima ir su duomenimis susijusios rizikos valdymą: Duomenys yra ne tik turtas, bet ir rizika organizacijai. Duomenys gali būti prarasti, pavogti arba netinkamai panaudoti. Organizacijos turi atsižvelgti į netinkamo duomenų naudojimo ir saugojimo pasekmes, o su duomenimis susijusi rizika turi būti valdoma kuo efektyviau su siekiu jos išvengti. (Data Management Body...) Vieni pagrindinių iššūkių, kurie sukelias rizikas yra: duomenų kiekis, kurį reikia apdoroti; duomenys neretai saugomi skirtingose repozitorijose; duomenų kokybės užtikrinimas; procesų trūkumas; duomenų integracijų trūkumas; specialistų trūkumas; duomenų kontrolės/valdymo trūkumas; duomenų saugos niuansai; duomenų struktūros pertvarkymas: iš nestruktūruotų į struktūruotus. (Malak, 2021)

Vadovybės įsitraukimas. Veiksmingam duomenų valdymui reikalingas vadovų įsitraukimas. Duomenų valdymas apima sudėtingą procesų rinkinį, kuris, kad būtų veiksmingas, reikalauja koordinavimo, bendradarbiavimo ir įsipareigojimo. Tam reikia ne tik vadybos įgūdžių, bet ir vizijos bei tikslo, kuriuos suteikia komanda ir vadovybė. (Data Management Body...)

Kadangi duomenų valdymas yra toks svarbus organizacijai, natūralu, kad kartu su svarba atsiranda ir tam tikri duomenų valdymo iššūkiai:

1. Duomenis lengva kopijuoti ir perkelti, tačiau juos nelengva atkurti, jei jie prarandami arba sunaikinami. Duomenys yra dinamiški ir gali būti naudojami įvairiais tikslais. Organizacijos duomenys yra unikalūs, todėl jei šie duomenys (klientų sąrašai, produktų atsargos ar pretenzijų istorija) būtų prarasti ar sunaikinti, jų pakeisti būtų neįmanoma arba labai brangu. Duomenys kartais vadinami organizacijos metaturtu ir dažnai yra apibrėžiami kaip organizacijos įžvalgų pagrindas.

2. Dar vienas iššūkis, su kuriuo susiduria organizacijos yra duomenų vertės nustatymas. Dažnai pasitaiko atvejai, kad duomenų vertė priklauso nuo duomenų konteksto. Tai yra, kad gali nutikti

taip, kad vieni duomenys yra vertingi vienai organizacijai, o kitos organizacijos atžvilgiu duomenų vadinti vertingais nebūtų tikslinga. Taip pat iškyla laikinumo problema – tai kas vertinga šiandien, nebūtinai bus vertinga rytoj. Vis dėlto organizacijoje tam tikros duomenų rūšys laikui bėgant gali būti nuolat vertingos. Pavyzdžiui: patikima informacija apie klientus. Laikui bėgant informacija apie klientus gali net tapti vertingesnė, nes kaupiasi vis daugiau su klientų veikla susijusių duomenų. Duomenų vertinimo procesas taip pat gali būti naudojamas kaip pokyčių valdymo priemonė.

3. Aukšta duomenų kokybė – vienas svarbiausių duomenų valdymo tikslų. Organizacijos tvarko savo duomenis, nes nori juos naudoti. Jei jos negali jais pasikliauti, kad patenkintų verslo poreikius, pastangos juos rinkti, saugoti, apsaugoti ir suteikti prieigą prie jų nueina perniek. Tačiau žmonės, kurie nori naudotis duomenimis, negali sau leisti nekreipti dėmesio į jų kokybę. Jie paprastai daro prielaidą, kad duomenys yra patikimi ir verti pasitikėjimo, kol neturi priežasties tuo abejoti.

4. Duomenis skirtingose organizacijos vietose valdo komandos, atsakingos už skirtingus duomenų gyvavimo ciklo etapus. Duomenų valdymui reikalingi projektavimo įgūdžiai, kad būtų galima planuoti sistemas. Techniniai įgūdžiai, kad būtų galima administruoti techninę įrangą ir kurti programinę įrangą. Duomenų analizės įgūdžiai, kad būtų galima suprasti klausimus ir problemas. Analitiniai įgūdžiai, kad būtų galima interpretuoti duomenis. Strateginis mąstymas, kad būtų galima įžvelgti galimybes aptarnauti klientus ir siekti tikslų. Iššūkis – sujungti visas šias komandas į vieną visumą, kad būtų galima tikslingai siekti organizacijos tikslų, išnaudojant stipriausias savo savybes.

5. Prastos kokybės duomenys (netikslūs, neišsamūs ar pasenę) akivaizdžiai kelia riziką, nes juose esanti informacija nėra teisinga. Tačiau duomenys taip pat kelia riziką, nes jie gali būti neteisingai suprasti ir panaudoti. Didžiausią vertę organizacijoms suteikia aukščiausios kokybės duomenys - prieinami, aktualūs, išsamūs, tikslūs, nuoseklūs, savalaikiai, tinkami naudoti, prasmingi ir suprantami. (Data Management Body...)

Apibendrinus, galima teigti, kad duomenų valdymo iššūkiai gana sutampa, remiantis skirtingų šaltinių pateikiama informacija. Duomenų kokybės, duomenų laikymo skirtingose vietose ir duomenų praradimo ir saugos problemos – vienos jų. Nors yra pateikiami skirtingi šių problemų sprendimai,

galima paminėti, kad duomenų tinklo architektūra, iš esmės padeda išspręsti šias problemas. Duomenų tinklo architektūra pasižymi decentralizuotu duomenų valdymu, kuris išskirstytas į atskirus domenus, kurie turi savo savininkus. Domenais remta architektūra padeda išspręsti duomenų kokybės ir laiko skirtingose vietose problemą, nes už savo domeną atsakingi asmenys, dirba tik su tam domenui priskirta informacija ir gali lengvai atsekti duomenų savininkus ir duomenų šaltinius. Sumažėja ir duomenų saugos rizika, nes priėjimą prie duomenų turi ribotas skaičius naudotojų.

1.1 Duomenų ir informacijos valdymo architektūros

Duomenys ir informacija – vienas kitą papildantys komponentai. Suteikę duomenims kontektsą gauname informaciją, o norint turėti informacija – reikalingi duomenys. Panašiai galima apibūdinti ir duomenų ir informacijos valdymo architektūras. Duomenų valdymo architektūra nusako duomenų kelią organizacijos viduje (duomenų surinkimas, saugojimas ir judėjimas). Tuo tarpu informacijos valdymo architektūra surenka duomenis iš skirtingų taškų ir suteikia jiems prasmę. (Lebenthal, 2018)

Galima apžvelgti duomenų ir informacijos valdymo architektūras iš arčiau ir kiek detaliau. Aukšto lygio duomenų valdymo architektūra turi aiškų procesą, kaip ir iš kur duomenys yra paimami ir koks yra jų galutinis taškas. Remiantis Craig‘o Stedman‘o (angl. Craig Stedman) straipsniu, aukšto lygio architektūrą sudaro 6 sluoksniai. Pirmasis sluoksnis yra duomenų šaltinių sluoksnis, kuriame sukaupti įvairūs duomenys iš skirtingų šaltinių. Iš pirmojo sluoksnio duomenys toliau keliauja į antrąjį, kuris pavadintas operacinių duomenų sluoksniu. Ten duomenys atsiduria reliacinėse duomenų bazėse, debesijos saugyklose, srauto sistemose ir panašiai. Iš antrojo sluoksnio duomenys toliau keliauja į trečiąjį, kuris pavadintas – duomenų įkrovimo ir sustojimo sluoksnis. Šiame sluoksnyje duomenys yra prafiltruojami, išvalomi ir suskirstomi į tam tikras grupes/tipus. Ketvirtasis sluoksnis pavadintas – duomenų integravimo ir transformavimo sluoksnis. Čia duomenys įprastai yra paruošiami eksportui į analitinį sluoksnį, naudojant ETL/ELT (*angl. Extract Transform Load ir Extract Load Transform*) bei realaus laiko integravimo priemones. Iš ketvirtojo sluoksnio, duomenys toliau keliauja į minėtąjį analitinį sluoksnį, kuriame duomenys pakraunami į duomenų žemėlapius (*angl. Data Marts*), duomenų saugyklą, duomenų ežerą ar kitas duomenų saugojimo vietas. Ir galiausiai duomenys nukeliauja į paskutinį verslo įžvalgos ir analitikos sluoksnį, kuriame duomenys yra vizualizuojami ir jų pagalba atliekami įvairūs raportai ir analitiniai procesai. (Stedman, 2021)

Tuo tarpu informacijos valdymo architektūra laviruoja tarp trijų komponentų: naudotojų, turinio ir konteksto. Tęsiant apie kiekvieną komponentą atskirai, kontekstas šioje trijų komponentų sudėtyje

apima: verslo tikslus, technologijas, išteklius, kultūrą bei įvairius apribojimus. Turinys apima: tikslus, dokumentų ir duomenų tipus, kiekį, struktūrą bei duomenų valdymo įgalinimą ir nuosavybės teisę. Ir galiausiai naudotojai: tikslinė rinka, poreikiai, užduotys ir patirtis. (Information Architecture Basics) Realiai informacijos valdymo architektūrą galima susieti ir su duomenų valdymo architektūra, nes duomenų valdymo architektūroje apžvelgus skirtingus sluoksnius buvo sužinota, kad po duomenų paruošimo eksportui jie keliauja į analitinį sluoksnį. Būtent nuo analitinio sluoksnio galėtų įsijungti informacijos valdymo architektūra, nes šioje stadijoje duomenys jau turi kontekstą, jie turi prasmę ir yra paruošti naudoti.

Apžvelgus duomenų valdymo architektūros ir informacijos valdymo architektūros ypatumus, galima šias architektūras apibendrinti. Duomenų valdymo architektūra apibrėžia kaip duomenys bus surenkami ir organizuojami. Įprastai sukuriamos taisyklės ir praktikos, kad būtų paprasčiau sekti duomenų kelią proceso metu ir palengvinti jų apdorojimą. Kai duomenys iš „padrikų“ tampa struktūruotais ir sutvarkytais, jie yra perimami informacijos valdymo architektūros. Ir galiausiai, informacijos valdymo architektūra pasinaudodama duomenimis gali teikti konkrečias ir tikslingas įžvalgas. Šių architektūrų pagalba gauname patikimą informaciją ir kokybiškus duomenis.

1.2 Metaduomenų valdymas šiuolaikinėse organizacijose

Siekiant remti įvairius organizacinius vaidmenis, kaip operacijas, stebėseną, kontrolę ir sprendimų priėmimą, naudojamos kompiuterinės duomenų apdorojimo ir informacijos sistemos. Organizacijos turi daugybę tokių sistemų, kurių kiekviena turi konkrečią apimtį, paskirtį ir funkcionalumą. Duomenų analizės ir sprendimų palaikymo sistemos įvertina verslo rezultatus pagal užsibrėžtų tikslų rinkinį ir padeda priimti naujus verslo sprendimus. Duomenų saugyklos technologija suteikia įrankius ir metodus, skirtus efektyviai organizuoti, naršyti, užklausti ir vizualizuoti didelį kiekį istorinių duomenų analizei ir sprendimų priėmimui.

Duomenų saugyklos duomenys paprastai organizuojami kaip „žvaigždžių schema“, „multi-dimensinio kubo schema“ arba reliacinis modelis. Operacinėse sistemose pateikti duomenys keliauja į ETL etapą, kai atitinkami duomenys išgaunami, transformuojami, susiję su bendra veikla ir galiausiai pakraunami į saugyklą. Didelėms organizacijoms šiandien reikia lanksčios prieigos prie įvairios informacijos, kuri yra jos operacinėse sistemose. Duomenų saugyklos technologija palengvina integruotų ir dalykinių istorijos duomenų kūrimą ir suteikia lanksčius būdus pasiekti, apibendrinti ir vizualizuoti informaciją. Kad naudotojai žinotų apie turimus duomenis ir jų struktūrą, reikia sukurti metaduomenų

saugyklas, kuriose būtų pateikti išsamūs duomenų aprašymai. Metaduomenys iš esmės suklasifikuoti į verslo ir techninius metaduomenis. Būtent verslo metaduomenys sudaro tinkamas verslo aplinkybes duomenų supratimui ir analizei ir sprendimų priėmimui. Atsižvelgiant į metaduomenų svarbą, pagrindiniai standartizacijos standartai yra pasiekti daugiausia dėmesio skiriant techniniams metaduomenims. Verslo metaduomenys kelia daug iššūkių, susijusių su apimtimi, abstrakcijomis ir struktūra. Vienas iš sprendimų gali būti integruotas metaduomenų ir duomenų saugyklos modelis. Modeliu siekiama integruoti tris segmentus: verslo metaduomenis, techninius metaduomenis ir duomenų saugyklą, kad naudotojai galėtų interpretuoti, suprasti, pasiekti ir analizuoti duomenų saugyklos duomenis, atsižvelgiant į metaduomenų ir duomenų laiko charakteristikas, o tada įvertinti verslo tikslų pasiekimą ir planuoti būsimus verslo veiksmus. Tai yra esminiai reikalavimai, kad būtų galima atlikti reguliavimo ir planavimo funkcijas. Tam pasiekti, naudojamas objektinis modeliavimas naudojant UML (*angl. Unified Modeling Language*) vaizdą. Apibrėžiamas visų klasių laikinumas ir navigaciniai metodai, kad būtų galima gauti visus susijusius metaduomenis ir bet kurio objekto duomenis. (Structuring Business Metadata...)

Rinkoje egzistuoja nemažai įrankių, kurie skirti padėti įmonėms valdyti metaduomenis. Populiarūs rinkos įrankiai naudoja tokias funkcijas kaip: metaduomenų integravimą kelioms duomenų bazių valdymo sistemų platformoms su prieinamais verslo įžvalgos platformų integratoriais, duomenų modeliavimo sprendimais ir kita duomenų integravimo programine įranga. Naudotojai taip pat gali gauti metaduomenis iš kitų priemonių, kad sukurtų duomenų perkėlimo ir duomenų vertinimo ataskaitų teikimo ir stebėsenos sistemas. Dar vienas aktualus funkcionalumas yra išsami metaduomenų analizė. Kai kurie produktai leidžia ieškoti ir naršyti metaduomenis, taip pat teikti duomenų kelio analizę, poveikio analizę, semantinį apibrėžimą ir semantinę bet kurio metaduomenų turto naudojimo analizę kataloge. Kai kurie įrankiai turi ir savo metaduomenų saugyklas, kuriose saugomi metaduomenys. Naudotojai turi galimybę importuoti metaduomenis į saugyklą iš kelių šaltinių, eksportuoti metaduomenis įvairiais metodais ir perkelti metaduomenis iš skirtingų saugyklų. Saugykloje atliekami pakeitimai automatiškai atliekami visame pakete ir naudoja standartinę sąryšinės duomenų bazės technologiją. (The best metadata...)

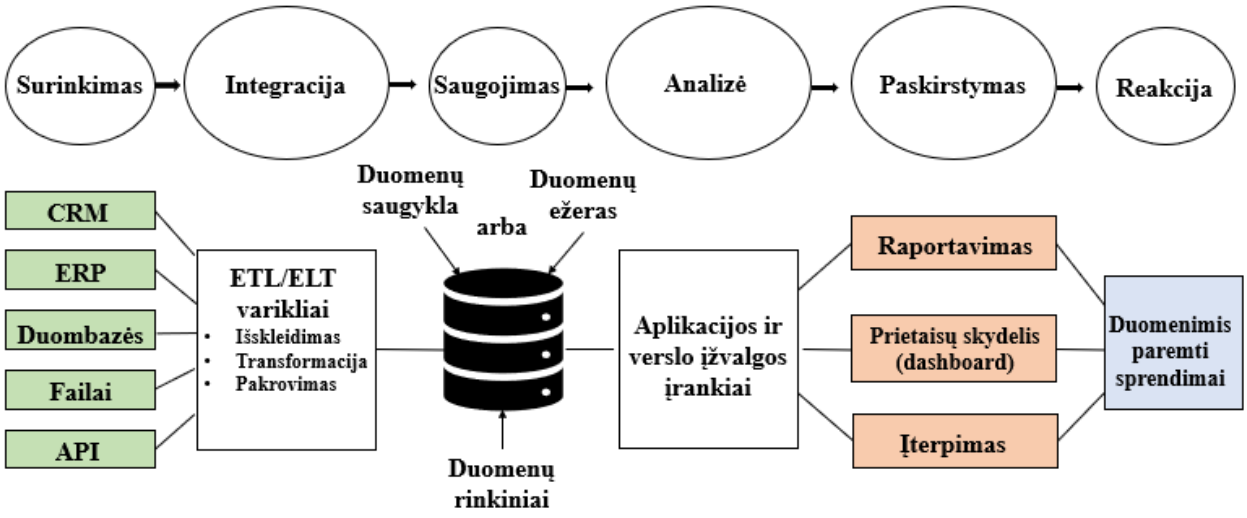
2. Klasikinė verslo įžvalgos architektūra ir analitinių duomenų repozitorijos

Verslo įžvalgos architektūra yra terminas, kuris naudojamas apibūdinti duomenų tvarkymo standartams ir strategijoms, naudojant kompiuterinius metodus ir technologijas, sukuriančias verslo įžvalgos sistemas. Vienas iš verslo įžvalgos architektūros komponentų yra duomenų saugojimas. Duomenų saugojimą, valymą ir išgavimą turi vykdyti centrinė saugyklų sistema, būtent todėl duomenų saugykla yra laikoma pagrindine verslo įžvalgos sudedamąja dalimi.

Tvirtą verslo įžvalgos sistemą sudaro šie komponentai:

- Duomenų rinkimas
- Duomenų integracija
- Duomenų saugojimas
- Duomenų analizė
- Duomenų platinimas
- Reakcija, pagrįsta įžvalgomis

Šioje diagramoje galima matyti visą proceso seką, tam tikrais sluoksniais. Šiuo atveju, duomenų rinkimas atitinka duomenų rinkimą, duomenų integracija atitinka duomenų ruošimą/rengimą ir duomenų įvestį, duomenų saugojimas atitinka duomenų talpinimą į saugyklą, duomenų analizė atitinka apdorojimą, o duomenų platinimas atitinka duomenų išvestį. Žemiau yra išskiriami verslo įžvalgos sistemos komponentai:



1 pav. *Verslo įžvalgos architektūra (šaltinis: sukurta autoriaus, remiantis B.Calzon (2022))*

Duomenų surinkimas. Pirmasis stabilios architektūros kūrimo žingsnis prasideda renkant duomenis iš įvairių šaltinių, tokių kaip: CRM (*angl. Customer Relationship Management*), ERP (*angl. Enterprise Resource Planning*), duomenų bazių, failų ar API (*angl. Application Programming Interface*). Tai vyksta atsižvelgiant į įmonės reikalavimus ir išteklius. Šiuolaikiški verslo įžvalgos įrankiai siūlo daugybę skirtingų, greitų ir lengvų duomenų jungčių, kad šis procesas vyktų sklandžiai ir lengvai, naudojant fone esančius ETL ir ELT variklius. Šie du varikliai skiriasi savo procesa eiga: ETL procesas duomenis išgauna, transformuoja ir užkrauna, tuo tarpu ELT išgauna, užkrauna ir tada transformuoja. Duomenų saugykla įprastai naudoja ETL technologiją, o duomenų ežeras ELT. Varikliai įgalina ryšį tarp išsibarsčiusių skyrių ir sistemų, kurios kitu atveju liktų skirtingos. Verslo požiūriu tai yra esminis elementas, kuriant sėkmingą, į duomenis orientuotą sistemą, kuri gali pašalinti klaidas ir padidinti našumą.

Duomenų integracija. Kai duomenys surenkami iš skirtingų šaltinių, sekantis žingsnis yra juos suintegruoti ir įkelti į duomenų saugyklą arba įkelti pradinės formos į duomenų ežerą. Vis didėjant duomenų kiekiui ir taip apkraunant informacinių technologijų departamentus bei specialistus, ETL ir ELT kaip paslauga yra natūralus atsakymas, siekiant išspręsti sudėtingas duomenų užklausas įvairiose pramonės šakose.

Duomenų saugojimas. Į duomenų saugyklą arba duomenų ežerą sukelti duomenys yra saugomi ir paruošti naudojimui.

Duomenų analizė. Visuotinis sėkmingos analizės poreikis, leidžiantis bet kokio dydžio įmonėms augti ir gauti pelną yra atliekama naudojant verslo įžvalgos įrankius. Ypač, kai kalbama apie Ad Hoc analizę, kuri suteikia laisvę, patogumą ir lankstumą atliekant analizę ir padedant greitai ir tiksliai atsakyti į svarbius verslo klausimus. Trumpai tariant, Ad Hoc analizė yra verslo įžvalgos procesas, skirtas atsakyti į vieną konkretų verslo klausimą.

Duomenų platinimas/paskirstymas. Duomenų platinimas yra vienas iš svarbiausių procesų, kai reikia dalintis informacija ir suteikti suinteresuotoms šalims būtinas įžvalgas siekiant tvarios verslo plėtros. Duomenis galima perduoti arba suteikti prieigą prie jų. Kaip pavyzdį galima paimti įmones, kurios suteikia savo analitikams prieigą prie duomenų saugyklos ar duomenų ežero ir išteklių, tačiau neleidžia tų duomenų keisti. Tiesa duomenų ežere duomenys gali būti pakoreguoti.

Reakcija, pagrįsta įžvalgomis. Tai paskutinis etapas, kuomet įmonė turi pradėti priimti sprendimus, remdamasi duomenimis ir gautomis įžvalgomis. Be duomenų ir jų saugojimo šis etapas nebūtų įmanomas ir verslas negalėtų progresuoti. Kaip pavyzdį galima paimti KPI (*angl. Key Performance Indicator*). Pagrindiniai veiklos rodikliai (KPI) – tai tam tikri matuokliai, matuojantys svarbiausius verslo veiklos objektus. Tam, kad KPI galėtų išmatuoti verslo veiklos objektus yra reikalingi duomenų saugykloje esantys struktūrizuoti duomenys. KPI pagalba yra išmatuojama vertė, parodanti, kaip efektyviai įmonė siekia pagrindinių verslo tikslų. Duomenų ežero naudojimo atveju, duomenys visų pirma turėtų būti paruošti naudojimui ir tik tada naudojami vertei išmatuoti. (Leščevskij, 2019)(Calzon, 2022)

Analitiniai duomenys dažniausiai yra tokie duomenys, kurie padeda organizacijoms priimti įžvalgomis paremtus sprendimus ir rengti įvarias ataskaitas. Tai yra istoriniai duomenys (ne realaus laiko duomenys), kurie įprastai saugomi skaitymui skirtose (*angl. schema-on-read*) duomenų repozitorijose, kurios optimizuotos duomenų analizei. (Spacey, 2018) Viena iš tokių repozitorijų, kuri analizuoja istorinius duomenis yra duomenų saugykla. Duomenų saugykla yra gana sena ir nusistovėjusi technologija. Nuo 1990 iki 2000 metų įvykęs technologinis ir kultūrinis sprogimas išaugino interneto populiarumą. Populiarėjant internetui ir atsirandant naujoms technologijoms, augo ir duomenų kiekis. Iki 2000 metų daugelis organizacijų pastebėjo, kad plečiantis duomenų bazėms ir taikomosioms sistemoms, jų sistemos buvo blogai integruotos, o duomenys buvo nenuoseklūs. Buvo pastebėta, kad gaunama ir saugoma daugybė fragmentuotų duomenų. Duomenis buvo reikalinga kažkaip integruoti, kad jie suteiktų svarbią verslui informaciją, kuri reikalinga priimant sprendimus konkurencingoje ir nuolat besikeičiančioje rinkoje. Tam ir buvo pradėtos kurti duomenų saugyklos, kad duomenys būtų konsoliduojami ir būtų suteikta pagalba, priimant strateginius sprendimus. (Foote, 2018) Bėgant metams susikūrė duomenų ežero architektūra, kuri buvo patraukli tuom, kad į duomenų ežerą galima kelti bet kokius duomenis. Priešingai, nei duomenų saugykloje, kur keliami tik struktūruoti duomenys, duomenų ežeras priima įvairaus formato duomenis ir tik paskui juos transformuoja naudojant ELT procesą. Duomenų ežero idėja buvo patraukli ir todėl, nes prasidėjus didžiųjų duomenų erai, duomenų kiekis augo labai dideliu greičiu, o duomenų ežero ypatumas kelti bet kokius duomenis, atrodė kaip visai patrauklus sprendimas sutaupyti laiko ir resursų. Tačiau laikui bėgant pradėjo ryškėti ir duomenų ežero spragos: nepatogus duomenų sujungimas, sunkus duomenų savininkų ir duomenų šaltinių atsekamumas, integracijų problemos ir panašiai. Šias ir kitas problemas taikosi spręsti naujai besikurianti duomenų tinklo architekūra. Apie šią, o taip pat ir duomenų saugyklos ir duomenų ežero architektūrą, jų panašumus ir skirtumus, bus rašoma sekančiuose skyriuose.

2.1 Klasikinė duomenų saugykla

Duomenų saugykla yra sistema, kuri periodiškai nuskaito ir sujungia duomenis iš skirtingų šaltinių į normalizuotą duomenų talpyklą. Paprastai ten saugoma daugelio metų ir didelio mąsto informacija, kuri yra užklausiama verslo įžvalgos pagalba arba atliekant analizes. Duomenų saugyklos yra atnaujinamos periodiškai, o ne kiekvieną kartą atlikus tam tikrą operaciją. (Oketunji ir Omodara, 2011). Nepaisant to, kad duomenų saugykla yra gana sena ir jau nusistovėjusi technologija, ji toliau yra plačiai naudojama daugelyje šiuolaikinių architektūrų. Todėl yra įdomu panagrinėti duomenų saugyklų technologinius įgalintojus, tokius kaip: masinis paralelinis apdorojimas, ETL ir ELT procesai, atminties technologija, SQL (*angl. Structured Query Language*) palaikymas, „debesijos“ sprendimai ir skirtingi duomenų modeliai. Šie įgalintojai leidžia šiai technologijai išlikti ir būti plačiai naudojamai.

Duomenų saugyklos informacijos šaltinis yra operatyvinė duomenų bazė. Duomenų saugyklų sistemos dar vadinamos OLAP (*angl. Online-Analytical Processing*), o operatyvinės duomenų bazės sistemos OLTP (*angl. Online Transactions Processing Databases*). (Operational Database vs…) Štai kelios priežastys, dėl kurių duomenų saugyklos ir operatyvinės duomenų bazės laikomos atskirai:

1. Operatyvinė duomenų bazė yra sukuriama konkrečioms užduotims ir tokioms apkrovoms kaip: paieška, indeksavimas ir t.t. Duomenų saugyklose dažniausiai yra pateikiamos sudėtingos užklausos, o jose pateikiama bendra duomenų forma.
2. Operatyvinės duomenų bazės palaiko kelių operacijų apdorojimą vienu metu. Veikimui reikalingi kontrolės lygiavertiškumo ir atkūrimo mechanizmai, kad būtų užtikrinas duomenų bazės patikimumas ir nuoseklumas.
3. Operatyvinės duomenų bazės užklausa leidžia nuskaityti ir modifikuoti operacijas, o duomenų saugyklos užklausai reikia tik skaityti prieigą prie saugomų duomenų.
4. Operatyvinėje duomenų bazėje saugomi dabartiniai duomenys, o duomenų saugykloje yra saugomi istoriniai (*angl. Historical data*) duomenys. (Data Warehousing Tutorial...)

Duomenų saugyklos bruožai:

- **Dalykiškai orientuota** – duomenys iš įvairių šaltinių yra suskirstyti į grupes apie bendras temas, į kurias organizacija norėtų atkreipti dėmesį. Tai gali būti: klientai, pardavimai, produktai ir t.t.

- **Integruota** – duomenų saugykla yra sukuriama integruojant duomenis iš heterogeninių šaltinių, tokių kaip realicinės duomenų bazės, paprastos rinkmenos ir t.t. Ši integracija pagerina duomenų analizės efektyvumą.
- **Kintanti laike** – duomenų saugykloje surinkti duomenys identifikuojami tam tikrais laikotarpiais. Duomenys, duomenų saugykloje teikia informaciją iš istorinės pusės.
- **Pastovi** – pridėti naujus duomenis į duomenų saugyklą galima reguliariai. Bet saugykloje saugomi duomenys yra tik skaitymui. Tai reiškia, kad naudotojui nėra leidžiama atnaujinti, perrašyti ar ištrinti saugomų duomenų.
  (Data Warehousing Tutorial...)

Duomenų saugyklos analizės ir apdorojimo tipai:

1. Informacijos apdorojimas – duomenų saugykla leidžia apdoroti joje saugomus duomenis. Jie gali būti apdorojami užklausomis, pagrindinėmis statstinėmis analizėmis arba ataskaitomis, naudojant kryžmines lenteles, lenteles, diagramas ar schemas.
2. Analitinis apdorojimas – duomenų saugykla palaiko analitinį apdorojimą su joje esančia informacija. Duomenys gali būti analizuojami naudojant pagrindines OLAP operacijas, įskaitant ir paprasčiausias OLAP operacijas: pjaustymas (*angl. slice and dice*), suskaidymas (*angl. drill-down*), agregavimas/apibendrinimas (*angl. roll-up*), ir sukimas (*angl. pivoting*).
3. Duomenų gavyba – duomenų gavyba padeda atrasti žinias, ieškant paslėptų modelių ir asociacijų, konstruojant analitinius modelius, atliekant klasifikavimus ir įžvalgas. Šie kasybos rezultatai gali būti pateikti naudojant vizualizacijos priemones. (Data Warehousing Tutorial...)

2.2 Duomenų saugyklos architektūra

Duomenų saugyklos architektūrą galima pavadinti metodu, kuriuo apibrėžiama bendra duomenų perdavimo bei apdorojimo ir pateikimo architektūra, skirta galutiniams naudotojams, dirbantiems organizacijoje. Kiekviena duomenų saugykla yra skirtinga, tačiau visoms būdingi standartiniai esminiai komponentai. (Data Warehouse Architecture...) Ji skirta užklausoms ir analizei, o joje paprastai pateikiami istoriniai duomenys. Duomenų saugyklos atskiria analizės darbo krūvį nuo sandorių darbo krūvio ir leidžia organizacijai konsoliduoti duomenis iš kelių šaltinių. Tai padeda išlaikyti istorinius

įrašus ir analizuoti duomenis, kad būtų galima geriau suprasti verslą ir jį tobulinti. Be reliacinės duomenų bazės, duomenų saugyklos aplinkoje egzistuoja ETL technologija, statistinės analizės, ataskaitų rengimo bei duomenų gavybos galimybės. Taip pat klientų analizės įrankiai ir kitos programos, kurios valdo duomenų rinkimo, transformavimo į naudingą ir praktiškai pritaikomą informaciją bei pateikimo verslo naudotojams procesą. (What Is a...)

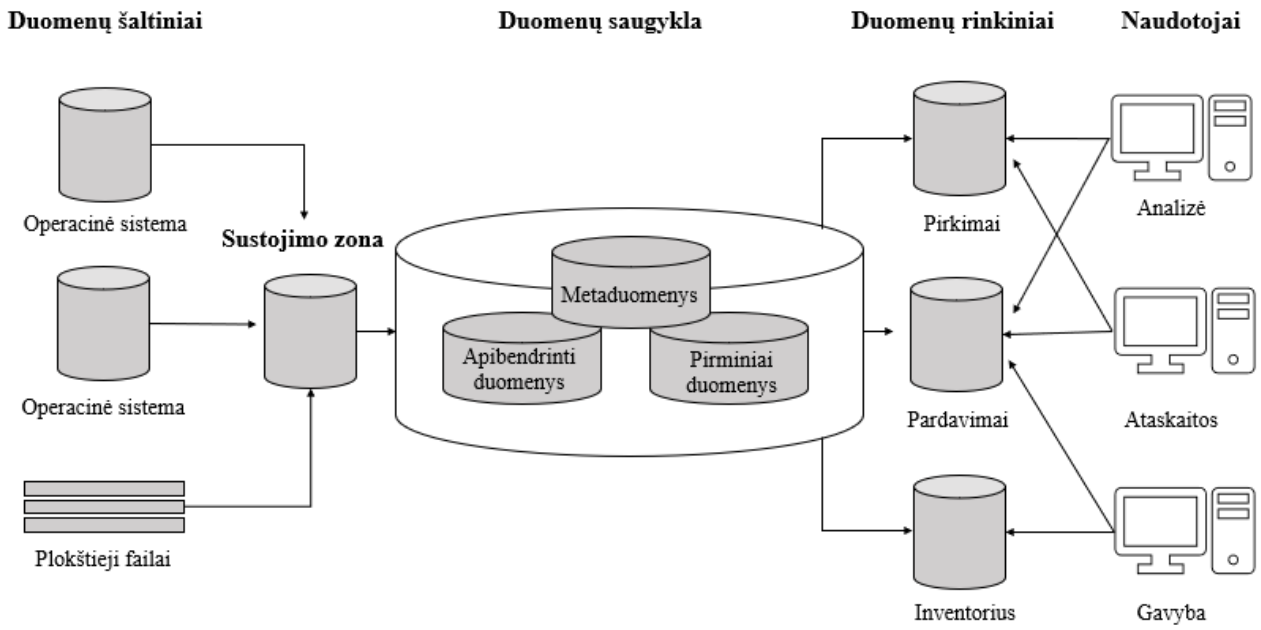
Kad būtų pasiektas kokybiškesnis verslo įžvalgos rezultatas, duomenų saugykloje dirbama su duomenimis, surinktais iš įvairių šaltinių. Duomenų šaltiniai gali būti gaunami iš viduje sukurtų sistemų, įsigytų programų, trečiųjų šalių duomenų ir kitų šaltinių. Jie gali būti susiję su sandoriais, gamyba, rinkodara, žmogiškaisiais ištekliais ir kitais. Duomenų saugykloje paprastai saugomi daugelio mėnesių ar metų duomenys, kad būtų galima atlikti istorinę analizę. Duomenys į duomenų saugyklą paprastai įkeliami naudojant ETL procesą iš daugelio duomenų šaltinių. Šiuolaikinės duomenų saugyklos naudoja ir ELT variklį, kai visa arba didžioji dalis duomenų transformavimo atliekama duomenų bazėje, kurioje yra duomenų saugykla. ETL operacijų greitis ir patikimumas yra duomenų saugyklos pagrindas, nes šis procesas stipriai įtakoja technologijos našumą. Duomenų saugyklos naudotojai atlieka duomenų analizę, kuri dažnai susijusi su laiku: pavyzdžiui, praėjusių metų pardavimų duomenų konsolidavimas, atsargų analizė, pelno pagal produktus ir klientus analizė ir kitos panašios užklausos. Tačiau naudotojai, orientuoti į laiką ar ne, nori nagrinėti duomenis taip, kaip jiems atrodo tinkama, o gerai suprojektuota duomenų saugykla turėtų būti pakankamai lanksti, kad atitiktų šiuos reikalavimus. Kartais naudotojams reikalingi labai apibendrinti duomenys, o kartais jiems reikia detalizuotų duomenų. Sudėtingesnės analizės apima tendencijų analizę ir duomenų gavybą, kai esami duomenys naudojami tendencijoms prognozuoti arba ateities įžvalgoms prognozuoti. Duomenų saugykla yra pagrindinis variklis, naudojamas tarpinės programinės įrangos verslo žvalgybos aplinkose, kuriose galutiniams naudotojams pateikiamos ataskaitos, skydeliai ir kitos sąsajos. (What Is a...)

Duomenų rinkiniai atlieka tą pačią funkciją kaip ir duomenų saugykla, tačiau jų taikymo sritis yra sąmoningai ribota. Duomenų rinkiniai gali būti skirti vienam konkrečiam skyriui ar verslo linijai. (Data Warehouse vs...) Duomenų rinkinio privalumas, palyginti su duomenų saugykla, yra tas, kad dėl ribotos aprėpties jį galima sukurti daug greičiau. Tačiau duomenų rinkiniuose taip pat kyla problemų dėl nenuoseklumo. Reikia griežtos drausmės, kad duomenys ir skaičiavimų apibrėžtys būtų nuoseklūs visose duomenų rinkiniuose. Ši problema buvo plačiai pripažinta, todėl duomenų rinkiniai egzistuoja dviejų stilių. Nepriklausomi duomenų rinkiniai - tai tokie rinkiniai, kurie yra maitinami tiesiogiai iš šaltinio duomenų. Priklausomi duomenų rinkiniai yra maitinami iš esamos duomenų saugyklos. Priklausomi duomenų rinkiniai, priešingai nei nepriklausomi, gali padėti išvengti nenuoseklumo problemų, tačiau

jiems reikia, kad jau egzistuotų įmonės lygmens duomenų saugykla. (What Is a...) Todėl buvo pasirinkta pavaizduoti duomenų saugyklos architektūra su sustojimo zonomis (*angl. staging area*), kuriose duomenys yra išvalomi ir duomenų rinkiniais, kurie teikia ir ima duomenis iš duomenų saugyklos.

Duomenų saugyklos architektūra su sustojimo zonomis ir duomenų rinkiniais (žr. 4 pav.) sudaryta iš 5 dalių: duomenų šaltinių, sustojimo zonos, duomenų saugyklos, duomenų rinkinių ir naudotojų. Iš duomenų šaltinių, t.y. operacinių sistemų ir plokščiųjų failų atkeliauja duomenys, kurie nugula sustojimo zonoje. Tai supaprastina duomenų išvalymą ir konsolidavimą. Atsidūrę duomenų saugykloje duomenys yra suskirstomi į tris grupes: metaduomenis, apibendrintus duomenis ir pirminius duomenis. Metaduomenys ir pirminiai duomenys yra standartiniai OLTP sistemos duomenys, o apibendrinti duomenys padeda duomenų saugyklai prasukti ilgai besisukančias užklausas. Toliau norint pritaikyti saugyklos architektūrą kelioms organizacijoms grupėms, tai galima padaryti sukuriant duomenų rinkinius. Duomenų rinkiniai yra duomenų saugyklos segmentas, kuriame galima pateikti informaciją ataskaitai ir analizuoti skyriaus, padalinio ar organizacijos veiklą.

Kaip šiuo pavyzdžiu, duomenų rinkiniai yra išskirti į pirkimų, pardavimų ir inventoriaus rinkinius. Tai tokiu atveju naudotojas norėdamas išanalizuoti istorinius duomenis apie pardavimus ar pirkimus galės pasinauoti pirkimų ar pardavimų duomenų rinkiniais. Taip pat gali būti poreikis gauti įžvalgą ar prognozę sprendimų priėmimui. Tokiu būdu, naudotojui suteikiama galimybė pasinaudoti gavybos funkcija ir analizuot visus duomenų rinkinius.



2 pav. *Duomenų saugyklos architektūra su sustojimo zonomis ir duomenų rinkiniais (šaltinis: sukurta autoriaus remiantis (Database Data Warehouse...))*

2.3 Duomenų saugyklos „debesyje“

Paprastai tariant, debesų kompiuterija (*angl. Cloud computing*) reiškia duomenų ir programų saugojimą ir prieigą prie jų internetu, o ne kompiuterio kietajame diske (What is cloud…). Debesų duomenų saugykla yra pagrindinis bet kurios duomenų analizės infrastruktūros elementas, kuriame debesies tiekėjas skiria vietą, reikalingą jūsų duomenims saugoti ir skaičiavimo galią, reikalingą jiems apdoroti. (Bekker, 2019)

Pasataraisiais metais duomenų saugyklos persikelia į debesį. Naujosios duomenų saugyklos debesyje, neatitinka tradicinės architektūros, o kiekvienas tokios duomenų saugyklos pasiūlymas turi unikalią architektūrą. (Data Warehouse Cloud…) Šiais laikais debesų kompiuterija ir saugojimas sulaukia vis daugiau mokslininkų ir naudotojų dėmesio. Ji skirta dinamiškai keičiamo mastelio ištekliams teikti, pašalinant investicijų ir priežiūros sritį. Daugelis komercinių produktų iš „Amazon“, „Microsoft“, „EMC“ ir „IBM“ teikia ekonomiškai efektyvius sprendimus, tokius kaip valandinis, mėnesinis ar metinis atsiskaitymas. Debesija pateikia aplinką, kurioje galutinis naudotojas gali atlikti užduotis taip, tarsi duomenys būtų saugomi vietoje, kai jie iš tikrųjų saugomi nuotolinėse sistemose. Tokiai aplinkai sukurti reikia galingų kompiuterių, greito užduočių vykdymo strategijų ir didelės spartos ryšių tinklo. Debesijos architektūroje, atsižvelgiant į išteklių, pavyzdžiui, taikymo pobūdžio, saugojimo vietos ir procesoriaus ciklų prieinamumą, duomenys paskirstomi skirtingiems mazgams. Be to, fizinė duomenų vieta gali dinamiškai pasikeisti iš vieno mazgo į kitą. Tradicinėse paskirstytose duomenų bazėse, siekiant sumažinti mazgo ryšį užklausos metu, lentelės horizontaliai suskirstomos ant sujungimo atributų, o susijusios dalys saugomos toje pačioje fizinėje sistemoje. Debesies aplinkoje neįmanoma užtikrinti, kad šie susiję skaidiniai visada būtų saugomi toje pačioje fizinėje sistemoje. Taigi užklausų vykdymas debesies duomenų saugyklose tampa sudėtingesnis, kai užklausose yra įvairių lentelių skaidinių, saugomų skirtinguose mazguose. Šie sujungia poreikį ir ryšį tarp užklausos vykdymo mazgų, kad rastų teisingą rezultatą. Šis spartus mazgų bendravimas turės neigiamos įtakos užklausos veikimui ir padidina tinklo srautą. (Optimizing Communication for...)

Keletas „debesų“ pagrindu sukurtų duomenų saugyklų pavyzdžių:

- „Amazon Redshift“
- „Google BigQuery“

„Amazon Redshift“:

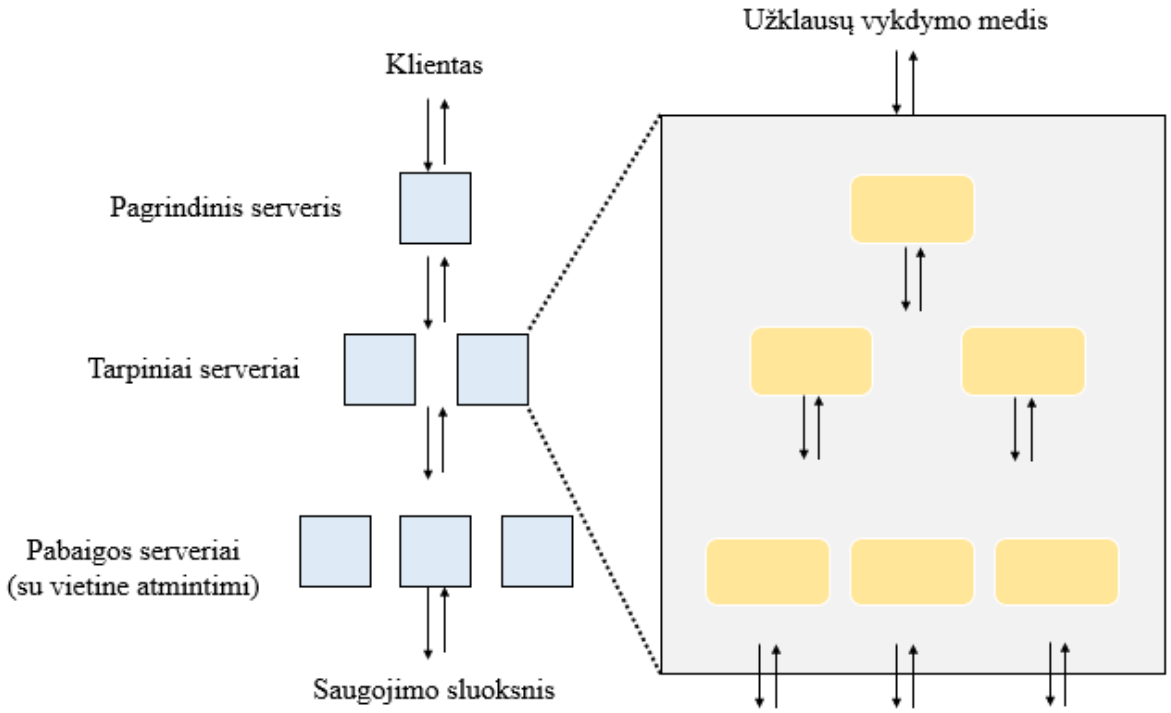
3 pav. *„Amazon Redshift” saugyklos architektūra. (šaltinis: sukurta autoriaus remiantis (Data Warehouse System...))*

„Amazon Redshift“ – saugykla debesyje, kuri yra paremta tradicinės duomenų saugyklos architektūros vaizdu.

Ši saugykla reikalauja, kad kompiuteriniai ištekliai būtų aprūpinti ir nustatyti klasterių pavidalu, kuriuose yra vieno ar daugiau mazgų rinkinys. Kiekvienas mazgas turi savo CPU (*angl. Central Processing Unit*), saugyklą ir operatyviąją atmintį (*RAM – Random Access Memory*). Lyderio mazgas kaupia užklausas ir perduoda jas skaičiavimo mazgams, kurie vykdo užklausas. Kiekvieno mazgo duomenys yra saugomi dalimis, vadinamuosiose pjūviuose (angl. slices). „Redshift“ naudoja stulpelio saugyklą, o tai reiškia, kad kiekviename duomenų bloke yra vertės iš vieno stulpelio keliose eilutėse, o ne viena eilutė su vertėmis iš kelių stulpelių. Čia naudojama MPP (*angl. Massively Parallel Processing*)

architektūra, kuri suskaido didelius duomenų rinkinius į dalis, kurios priskiriamos kiekvieno mazgo pjūviams. Užklausos vykdomos greičiau, nes skaičiavimo mazgai apdoroja užklausas kiekvienoje skiltyje vienu metu. Lyderio mazgas apibendrina rezultatus ir grąžina juos į kliento programą. Verta paminėti, kad „Redshift“ gali įkelti tik struktūrizuotus duomenis. Duomenis į „Redshift“ galima įkelti naudojant iš anksto integruotas sistemas („Amazon S3”, “DynamoDB”), perkeliant duomenis iš bet kokio pagrindinio kompiuterio, turinčio SSH (*angl. Shell Secure*) ryšį arba integruojant kitus duomenų šaltinius.

„Google BigQuery“:



4 pav. *„Google BigQuery“ saugyklos architektūra. (šaltinis: sukurta autoriaus remiantis (What Is BigQuery...)*

„BigQuery“ architektūra neturi serverio, o tai reiškia, kad „Google“ dinamiškai valdo mašinų išteklių paskirstymą. Todėl visi išteklių valdymo sprendimai yra paslėpti nuo naudotojo. „BigQuery“ leidžia klientams įkelti duomenis iš „Google Cloud Storage“ (tai „Google“ siūloma duomenų saugykla debesyje) ir kitų skaitomų šaltinių. Alternatyvus variantas yra duomenų srautas, o tai leidžia kūrėjams pridėti duomenis į duomenų saugyklą realiuoju laiku eilėmis po eilės, kai tik jie tampa prieinami. „BigQuery“ naudoja užklausos vykdymo variklį pavadinimu „Dremel“, kuris vos per kelias sekundes

gali nuskaityti milijardus duomenų eilučių. „Dremel“ naudoja daugybę lygiagrečių užklausų. Norėdamas nuskaityti duomenis pagrindinėje „Colossus“ failų valdymo sistemoje. „Colossus“ paskirsto failus 64 megabaitų dalimis tarp daugelio skaičiavimo išteklių, pavadintų mazgais, kurie yra sugrupuoti į grupes. „Dremel“ naudoja stulpelinę duomenų struktūrą, panašią į „Redshift“. Medžio architektūra per kelias sekundes išsiunčia užklausas iš tūkstančių mašinų.

Debesų saugyklos yra didelis žingsnis į priekį nuo tradicinės architektūros. Tačiau pereidami prie debesų saugyklų, naudotojai susiduria su tam tikrais iššūkiais:

- Duomenų įkėlimas į debesies duomenų saugyklą nėra nereikšmingas, o norint naudoti didelio masto duomenų perdavimus, reikia nustatyti, išbandyti ir peržiūrėti ETL procesą. Paprastai ši proceso dalis atliekama naudojant trečiųjų šalių įrankius.
- Atnaujinimai ir ištrynimai gali būti sudėtingi ir turi būti atliekami atsargiai, kad užklausos našumas nesumažėtų.
- Su pusiau struktūruotais duomenimis sunku susitvarkyti – juos reikia normalizuoti į reliacinį duomenų bazės formatą, kurį reikia automatizuoti dideliems duomenų srautams.
- Įdėtos struktūros paprastai nepalaikomos debesų duomenų saugyklose. Turėsite išlyginti įdėtas lenteles tokiu formatu, kokį supranta duomenų saugykla.
- Klasterio optimizavimas – yra įvairių variantų, kaip nustatyti klasterį, kad būtų galima vykdyti darbo krūvius. Skirtingam darbo krūviui, duomenų rinkiniams ar net skirtingo tipo užklausoms gali reikėti skirtingos sąrankos (*angl. setup*). Jei norite išlikti optimalūs, turėsite nuolat peržiūrėti ir taisyti savo sąranką.
- Užklausos optimizavimas – vartotojų užklausos gali neatitikti geriausios praktikos, todėl jų vykdymas užtruks daug ilgiau. Galima patiems dirbti su naudotojais ar automatinėmis klientų programomis, kad optimizuoti užklausas ir duomenų saugykla veiktų taip, kaip buvo tikėtasi.
- Atsarginė kopija ir atkūrimas – nors duomenų saugyklos tiekėjai teikia daugybę duomenų atsarginių kopijų atkūrimo galimybių, tačiau jų nustatymas nėra nereikšmingas, todėl juos reikia atidžiai stebėti. (Data Warehouse Cloud…)

2.4 Duomenų ežeras

Duomenys gali būti paimami iš įvairių šaltinių. Tuo tarpu duomenų valdymas, kaip jau ir minėta anksčiau šiame darbe – esminis faktorius siekiant kokybės ir efektyvumo, tačiau ir sudėtingas darbas. Organizacijos adaptuojasi prie duomenų ežero modelio, nes šis suteikia naudotojams neapdorotus ir

pirminius duomenis, kuriais gali būti eksperimentuojama ir atliekamos įvairios analizės. Duomenų ežeras gali tapti naujų ir istorinių duomenų sujungimo tašku, kuris leistų koreliuoti ir sujungti įvairius duomenis ir atlikti pažangias analizes naudojantis jais. Taip pat, duomenų ežero architektūra palaiko savitarnos duomenų praktiką, kas leidžia išgauti neatrastą vertę iš įvairių esamų ir naujų duomenų šaltinių. (Singh, 2019) Toliau šiame skyriuje bus aprašoma duomenų ežero architektūra, jos nauda ir iššūkiai bei tikslai.

Duomenų ežeras yra centralizuota duomenų saugykla/repozitorija, kuri gali saugoti didelį kiekį duomenų, pradedant struktūruotais ir baigiant pusiau struktūruotais arba visiškai nestruktūruotais duomenimis. Duomenų ežeras turi keičiamo dydžio saugyklą, kurioje galima tvarkyti didėjantį duomenų kiekį suteikiant duomenims judrumo ir pateikiant greitesnes įžvalgas. Šios architektūros savybė saugoti bet kokius duomenis, leidžia išsiskirti iš kitų tradicinių duomenų surinkimo ir saugojimo metodų bei pateikti naudotojui reikalingas įžvalgas. Duomenų ežeras turi plokščią architektūrą duomenims saugoti ir turi prieigą prie didelio kiekio duomenų ir informacijos pagal skaitymo schemą (*angl. schema-on-read*). (Singh, 2019) Įprastai ežeras talpinamas „Apache Hadoop“ sistemoje, kuri yra atviro kodo programinės įrangos sistema ir užtikrina patikimą didelių duomenų rinkinių apdorojimą, naudojant paprastus programavimo modelius. „Hadoop“ garsėja savo mastelio keitimo galimybėmis ir yra sukurta naudojant plataus naudojimo kompiuterių klasterius, kas leidžia priimti ekonomiškai efektyvius sprendimus, saugoti ir apdoroti didžiulius kiekius struktūruotų, pusiau struktūruotų ir nestruktūruotų duomenų. (What is Apache...)

Duomenų ežero architektūra susideda iš trijų sluoksnių ir trijų lygių. Įprastai sluoksniai tęsiasi per visus lygius, o lygiai yra juos papildantys komponentai. Sluoksnius sudaro: duomenų valdymo taisyklių įgalinimo ir saugos sluoksnis, metaduomenų sluoksnis ir informacijos gyvavimo ciklo valdymo sluoksnis. Lygius sudaro: duomenų priėmimo lygis, valdymo lygis ir paėmimo lygis. (Singh, 2019) Viena plačiai naudojama architektūra „Lambda“ puikiai apibūdina duomenų ežero architektūrą. „Lambda“ architektūra naudojama masiniams duomenų kiekiams apdoroti. (What is Lambda...) Pateikta kaip programinės įrangos projektavimo modelis, „Lambda“ architektūra sujungia internetinį ir paketinį (*angl. batch*) apdorojimą į vieną sistemą. Ji tinka programoms, kuriose naudojamas duomenų surinkimas, vėluoja duomenų prieinamumas (realaus laiko duomenų poreikis) ir reikalingas duomenų apdorojimas iškart jiems atkeliavus į sistemą. (Kiran, Murphy, Monga, Dugan ir Baveja, 2015) Būtent tai duomenų ežeras ir gali suteikti: duomenis galima skaityti realiu laiku, nes egzistuoja skaitymo schema bei duomenys galimi naudoti užkėlus juos į saugyklą. Tęsiant apie duomenų ežero architektūrą yra teigiama, kad pagal Brewer‘io teoremą, paskirstyta duomenų saugykla negali vienu metu teikti daugiau nei dviejų

iš trijų, teoremos išskirtų veiksnių: nuoseklumo, prieinamumo ir skirstymo tolerancijos. Todėl duomenų ežero architektūra, su „Lambdos“ architektūros pagalba, vadovaujasi šia Brewer‘io teorema. Dažniausiai prieinamumas pasirenkamas vietoje nuoseklumo, nes nuoseklumą galiausiai įmanoma pasiekti, tačiau be prieinamumo ir skirstymo tolerancijos saugykla praktiškai taptų neveiksni. Skirstymo tolerancija šiame kontekste reiškia, kad informacija ir duomenys yra paskirstyti tarp skirtingų sistemos mazgų, kurie atsakingi už sistemos veikimą. Todėl atsijungus vienam ar keliems mazgams, sistema nenustotų veikti, o tiesiog atsijungusių mazgų krūvis būtų perimtas kitų, veikiančių mazgų. (Singh, 2019)

Nepaisant visų privalumų, kiekviena architektūra turi ir savų iššūkių. Vieni tokių iššūkių, kalbant apie duomenų ežerą gali būti įvardinami kaip:

1. Reikalingas kokybiškas duomenų valdymas, nes be kokybiško ir efektyvaus duomenų valdymo, duomenų ežeras gali pavirsti į duomenų „pelkę“
2. Į duomenų ežerą gali patekti konfidencialumo reikalaujantys duomenys. Todėl tokiems duomenims reikalinga ypatinga priežiūra ir pamatuotas rizikos suvaldymas
3. Duomenų ežeras suteikia prieigą prie duomenų visiems organizacijos darbuotojams, tačiau ne visi darbuotojai turi pakankamai žinių suprasti nestruktūruotus arba pusiau struktūruotus duomenis, kurie reikalauja daugiau techninių žinių (Karatas, 2022)
4. Sudėtinga gauti įžvalgų iš buvusių darbuotojų, kurie dirbo su duomenims, nes nėra galimybės atsekti duomenų kelio ar duomenų šaltinio ir savininko
5. Saugumo problema. Duomenys gali būti pakeisti be papildomos peržiūros, nepaisant, kad gali būti tokiu duomenų, kuriems reikalinga papildoma saugos priežiūra. (Banafa, 2021)

3. Duomenų tinklo architektūra

Daugelis organizacijų investuoja į naujos kartos duomenų ežerą, su siekiu efektyvinti duomenų valdymą, kad būtų galima gauti verslo įžvalgų ir galiausiai priimti automatizuotus sprendimus. Tačiau duomenų ežero architektūra turi ir savų iššūkių: dažnu atveju nepatogu apjungti duomenis į vieną visumą, technologija priima skirtingų formatų duomenis, duomenų išgavimo sparta nėra didelė, duomenys turi neretai būti analizuojami iš naujo ir panašiai. Norėdami išspręsti šių nesėkmių problemas, turime pereiti nuo centralizuotos duomenų ežero ar jo pirmtako duomenų saugyklos idėjos. Reikia pereiti prie idėjos, kuri remiasi šiuolaikine paskirstyta architektūra: domenus laikyti pirmos klasės rūpesčiu, taikyti platforminį mąstymą, kad būtų sukurta savitarnos duomenų infrastruktūra, o duomenis traktuoti kaip produktą.

Tęsiant apie pačią duomenų tinklo architektūrą ir jos pradžią, galima paminėti, kad pirmoji, kuri apie tai pradėjo kalbėti ir vystyti šią idėja buvo Zhamak Dehghani. Zhamak Dehghani 2018 metais pradėjo skleisti žinią apie naująjį duomenų tinklo konceptą, kurio pagrindinis tikslas – duomenų decentralizacija, siekiant supaprastinti didžiųjų duomenų valdymą ir apskritai duomenų valdymą. (Zhamak Dehghani Director…)

Duomenų tinklo funkcijos ir savybės:

Zhamak Dehghani teigia, kad duomenų tinklas turėtų būti grindžiamas keturiais pagrindiniais principais:

1. Į domeną orientuota ir decentralizuota duomenų nuosavybė (*angl. domain-oriented and decentralized data ownership*)
2. Duomenys traktuojami kaip produktas (*angl. data as a product*)
3. Savitarnos duomenų platforma (*angl. self-serve data platform*)
4. Konsoliduotas duomenų valdymas (*angl. federated computational governance*)

Pirmoji duomenų tinklo funkcija susijusi su duomenų organizavimu atsižvelgiant į patį verslą. Žvelgiant į organizacinę struktūrą, suprantama, kad padaliniai apibrėžiami pagal veiklos sritis (logistikos, klientų aptarnavimo ir t.t.). Duomenų tinklo architektūra įgalina paskirstytos atsakomybės egzistavimą organizacijos komandoms, kurios gali geriau suprasti ir kurti savo konkrečios verslo srities duomenis. Tam, kad paskirstyti atsakomybę ir decentralizuoti jau žinomas monolitines architektūras, būtina modeliuoti dabartinę duomenų architektūrą, pagrįstą analitinių duomenų organizavimu pagal sritis. Kad

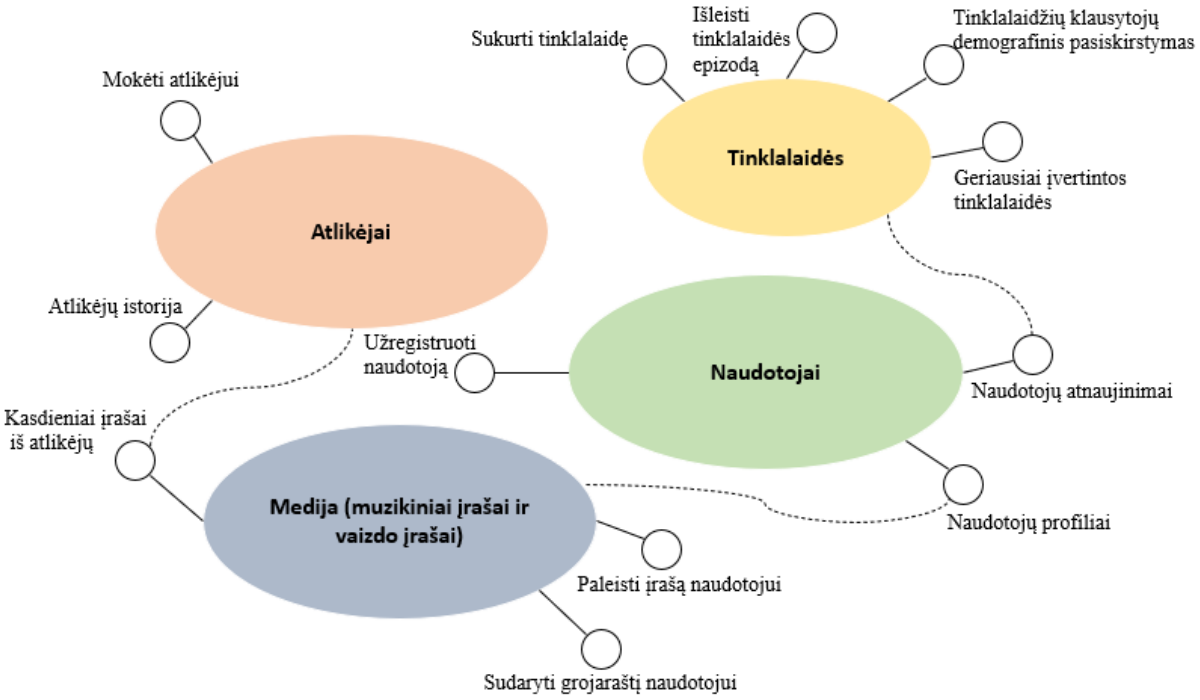
būtų galima geriau susidoroti su pokyčiais, domenai ir saugotų, ir teiktų savo duomenis. Sąnaudos, patiriamos ieškant kokybiškų ir aktualių duomenų yra viena iš problemų, su kuria susiduria kitų architektūrų naudotojai. Viena iš dažniausiai tokius sunkumus patiriančių architektūrų yra duomenų ežero architektūra. Dalis problemų kyla dėl to, kad net organizacijos, kurios laiko save orientuotomis į duomenis, su duomenimis nesielgia tinkamai. Duomenų demokratizacijos trūkumas yra viena iš netinkamo elgesio dalių. Todėl šioje vietoje iškyla antroji duomenų tinklo funkcija - duomenys turi būti traktuojami kaip produktas. Duomenų tinklo architektūra duomenims taiko jau žinomą „Product Thinking“ koncepciją, kad jie taptų svarbiausiu organizacijos prioritetu, o duomenų srautas ir saugojimas lieka antrame plane. Analitiniai duomenys dabar laikomi produktu, todėl turi būti užtikrinta jų kokybė, o šių duomenų vartotojai yra laikomi klientais, ir jų poreikiai turi būti tenkinami. Šis faktas reiškia, kad egzistuoja infrastruktūra, kuri leidžia komandoms kurti ir prižiūrėti savo duomenų produktus. Tačiau tam būtina, kad grupės turėtų prieigą prie aukšto lygio infrastruktūros, galinčios aprėpti visą jai būdingą sudėtingumą. Todėl šioje vietoje pasireiškia trečioji funkcija – savitarnos duomenų platforma, kuri įgalina komandas savarankiškai valdyti savo srities duomenis. Ir galiausiai kalbant apie ketvirtąją duomenų tinklo architektūros funkciją, t.y. konsoliduotą duomenų valdymą, turime kalbėti apie mechanizmą, kuris leistų sąveikauti tarp skirtingų domenų. (Dehghani, 2019)

Loginės architektūros pagal 4 pagrindinius duomenų tinklos architektūros principus:

Tam, kad būtų galima geriau įsivaizduoti duomenų tinklo architektūros veikimo principą, ji bus aiškinama per skaitmeninės medijos bendrovės panaudos atvejį. Galima įsivaizduoti, kad medijos bendrovė savo veiklą, taigi ir sistemas bei komandas, kurios palaiko šią veiklą, skirsto pagal sritis. Pavyzdžiui, „tinklalaidės“ - komandos ir sistemos, kurios valdo tinklalaidžių leidybą ir jų šeimininkus. „Atlikėjai“ - komandos ir sistemos, kurios valdo atlikėjų/kūrėjų priėmimą į platformą ir mokėjimą jiems. Duomenų tinklo idėja teigia, kad analitinių duomenų nuosavybė ir aptarnavimas turėtų atitikti šias sritis. Pavyzdžiui, komandos, kurios valdo tinklalaides, teikdamos aplikacijų programavimo sąsają (API) tinklalaidžių leidybai, taip pat turėtų būti atsakingos už istorinių duomenų teikimą, kurie atspindi „išleistas tinklalaides“ laike kartu su kitais faktais, tokiais kaip: „klausytojų skaičius“ laike.

Domenų nuosavybė (*angl. domain ownership*). Duomenų tinklas iš esmės grindžiamas decentralizavimu ir atsakomybės paskirstymu žmonėms, kurie yra arčiausiai duomenų. Natūraliai kyla klausimas, kaip išskaidyti ir decentralizuoti duomenų ekosistemos komponentus ir jų nuosavybę. Šiuo atveju komponentus sudaro analitiniai duomenys, jų metaduomenys ir jiems aptarnauti reikalingi skaičiavimai. Kad būtų galima skatinti tokį skaidymą, reikia sumodeliuoti architektūrą, kurioje analitiniai

duomenys būtų suskirstyti pagal sritis. Šioje architektūroje domeno sąsaja su likusia organizacijos dalimi apima ne tik veiklos galimybes, bet ir prieigą prie analitinių duomenų, kuriuos aptarnauja domenas. Pavyzdžiui, domenas „tinklalaidės“ teikia operacines API, kad būtų galima panaudoti funkciją „sukurti naują tinklalaidės epizodą“, bet taip pat ir analitinių duomenų galinį tašką, kad būtų galima gauti „visų tinklalaidžių epizodų duomenis per pastaruosius <n> mėnesius“. Tai reiškia, kad architektūra turi pašalinti bet kokią trintį, kad domenai galėtų teikti savo analitinius duomenis nepriklausomai nuo kitų domenų. Norint, išplėsti architektūros mastelį, ji turi palaikyti domenų komandų savarankiškumą išleidžiant ir diegiant jų operatyvinių ar analitinių duomenų sistemas. (Dehghani, 2019) Žemiau esantis pavyzdys 1 pav. iliustruoja atskirų domenų tarpusavio sąveikavimą.



5 pav. *Domenų atsakomybė už analitinius duomenis kartu su operaciniais gebėjimais (Šaltinis:sukurta autoriaus remiantis Zhamak Dehghani (2019))*

Šis pavyzdys parodo, kaip atskiri domenai, kurie yra atsakingi už savo srities duomenis, gali keistis informacija, kuri papildo vieną ar kitą domeną. „Atlikėjai“ siejasi su „Medijos (muzikos įrašai ir video įrašai) srautas“. „Medijos (muzikos įrašai ir video įrašai) srautas“ jungiasi su „Naudotojai“ ir galiausiai „Naudotojai“ siejasi su „Tinklalaidės“.

Duomenys traktuojami kaip produktas. Vienas iš esamų analitinių duomenų architektūrų iššūkių - didelė trintis ir išlaidos, susijusios su kokybiškų duomenų atradimu, supratimu, pasitikėjimu ir galiausiai naudojimu. Jei ši problema nesprendžiama, didėjant duomenų tinklui ji tik aštrėja, nes daugėja

vietų ir komandų, teikiančių duomenis - domenų. Duomenų kaip produkto principas skirtas duomenų kokybės ir senos duomenų „siloso“ problemos sprendimui. Arba kaip „Gartner“ naudoja tamsiųjų duomenų (*angl. dark data*) apibrėžimą tokiems duomenims: „Tamsieji duomenys – kaip informacijos turtas, kurį organizacijos renka, apdoroja ir saugo vykdydamos įprastą verslo veiklą, tačiau paprastai nenaudoja kitiems tikslams“. (Dark Data…) Analitiniai duomenys, kuriuos teikia sritys, turi būti traktuojami kaip produktas, o šių duomenų vartotojai turi būti traktuojami kaip klientai. Kartu su domenais atsiranda ir naujos rolės. Viena jų – domeno duomenų produkto savininkas, kurią turi įvesti organizacijos, atsakingos už objektyvias priemones, užtikrinančias, kad duomenys būtų pateikti kaip produktas. Šios priemonės apima duomenų kokybę, sutrumpėjusį duomenų vartojimo laiką ir apskritai duomenų naudotojų pasitenkinimą. Domeno duomenų produkto savininkas turi gerai suprasti, kas yra duomenų naudotojai, kaip jie naudoja duomenis ir kokie metodai jiems patogūs naudojant duomenis. Taip gerai pažįstant duomenų naudotojus, kuriamos jų poreikius atitinkančios duomenų produkto sąsajos. Visiems duomenų produktams galima sukurti standartizuotas sąsajas jiems palaikyti. Ryšys tarp duomenų naudotojų ir produktų savininkų yra būtina duomenų produktų sąsajų kūrimo dalis.

Žvelgiant iš architektūrinės pusės, siekiant palaikyti duomenis kaip produktą, kurį domenai gali savarankiškai teikti arba vartoti, duomenų tinklo architektūroje įvedama duomenų produkto sąvoka kaip architektūrinis kvantas.(Dehghani, 2019) Architektūrinis kvantas – yra mažiausias architektūros vienetas, kuris gali būti nepriklausomai įdiegtas ir pasižymėti dideliu funkciniu vientisumu bei apima visus funkcijai reikalingus struktūrinius elementus. (Ford, Parsons ir Kua, 2017) Siekiant atvaizduoti duomenų produkto kaip vieneto architektūrą, būtina aptarti tris struktūrinius komponentus, kurie yra reikalingi jo funkcijai atlikti, suteikiant prieigą prie domeno analitinių duomenų kaip produkto. Šie struktūriniai elementai yra:

- **Kodas** (*angl. code*) – a) Kodas, kuris skirtas duomenų srautams, atsakingiems už „prieš srovę“ (*angl. up-stream*) duomenų vartojimą, transformavimą ir teikimą. b) Kodas, skirtas API, suteikiančioms prieigą prie duomenų, semantinių ir sintaksės schemų, stebėjimo metrikų ir kitų metaduomenų. c) Kodas, skirtas tokiems požymiams, kaip prieigos kontrolės politika, atitiktis, kilmė ir t.t., užtikrinti.
- **Duomenys ir metaduomenys** – priklausomai nuo domeno duomenų pobūdžio ir jų vartojimo modelių, duomenys gali būti pateikiami kaip įvykiai, paketiniai failai, reliacinės lentelės, grafikai ir panašiai, išlaikant tą pačią semantiką. Kad duomenys būtų tinkami naudoti, yra susietas

metaduomenų rinkinys, įskaitant duomenų skaičiavimo dokumentaciją, semantinę ir sintaksės deklaraciją, kokybės metrikas ir panašiai.

- **Infrastruktūra** – infrastruktūros komponentas leidžia kurti, diegti ir paleisti duomenų produkto kodą, taip pat saugoti didžiuosius duomenis ir metaduomenis bei užtikrinti prieigą prie jų.

Savitarnos duomenų platforma. Galima įsivaizduoti, kad norint sukurti, įdiegti, vykdyti ir stebėti duomenų produktą - reikia sukurti ir paleisti nemažai infrastruktūros. Įgūdžiai, reikalingi šiai infrastruktūrai sukurti, yra specializuoti ir juos būtų sunku atkartoti kiekvienoje srityje. Svarbiausia, kad vienintelis būdas komandoms savarankiškai valdyti savo duomenų produktus yra turėti prieigą prie aukšto lygio infrastruktūros abstrakcijos, kuri pašalina sudėtingumą ir trintį, susijusią su duomenų produktų aprūpinimu ir jų gyvavimo ciklo valdymu. Todėl reikia naujo principo - savitarnos duomenų infrastruktūros, kaip platformos, kuri įgalina domenų autonomiją.

Tęsiant iš architektūrinės pusės, savitarnos duomenų platforma vaizduojama plokštumos principu. Savitarnos platformoje gali būti kelios plokštumos, kurių kiekviena skirta skirtingo profilio naudotojams. Toliau išskiriamos trys skirtingos duomenų platformos plokštumos:

1. **Duomenų infrastruktūros teikimo plokštuma** (*angl. Data infrastructure provisioning plane*) – ši plokštuma padeda užtikrinti pagrindinę infrastruktūrą, reikalingą duomenų produkto sudedamosioms dalims ir produktų tinklui veikti. Šią plokštumą sudaro: paskirstytosios failų saugyklos, prieigos kontrolės valdymo sistemos, orkestruotės, skirtos duomenų produktų vidiniam kodui paleisti ir t.t. Labiausiai tikėtina, kad šia sąsaja tiesiogiai naudosis arba kitų duomenų platformų plokštumos, arba tik pažengę duomenų produktų kūrėjai. Tai gana žemo lygio duomenų infrastruktūros gyvavimo ciklo valdymo plokštuma.
2. **Duomenų produkto kūrėjo patirties plokštuma** (*angl. Data product developer experience plane*) –tai yra pagrindinė sąsaja, kurią naudoja tipinis duomenų produktų kūrėjas. Ši sąsaja apima daugelį sudėtingų duomenų produktų kūrėjo darbo eigos palaikymo aspektų. Ji suteikia aukštesnį detalumo lygį nei anksčiau minėta „Duomenų infrastruktūros teikimo plokštuma“. Joje naudojamos paprastos deklaratyvios sąsajos duomenų produkto gyvavimo ciklui valdyti. Joje automatiškai įgyvendinamos tarpdalykinės problemos, apibrėžtos kaip standartų ir visuotinių susitarimų rinkinys, taikomas visiems duomenų produktams ir jų sąsajoms.

3. **Duomenų tinklo priežiūros plokštuma** (*angl. Data mesh supervision plane*) – Nors kiekvienos iš šių sąsajų įgyvendinimas gali remtis atskirų duomenų produktų galimybėmis, patogiau šias galimybes teikti duomenų tinklo priežiūros lygmeniu. Pavyzdžiui, gebėjimą atrasti duomenų produktus konkrečiam naudojimo atvejui geriausia užtikrinti atliekant paiešką arba naršant duomenų produktų tinkle. Norint susieti kelis duomenų produktus, kad būtų galima sukurti aukštesnio lygio įžvalgą, geriausia užtikrinti vykdant duomenų semantinę užklausą, kuri gali veikti keliuose duomenų tinkle esančiuose duomenų produktuose.

Konsoliduotas duomenų valdymas. Duomenų tinklo architektūra atitinka paskirstytos sistemos architektūrą. Tai nepriklausomų duomenų produktų rinkinys su nepriklausomu gyvavimo ciklu, kurį sukūrė ir įdiegė greičiausiai nepriklausomos komandos. Tačiau daugumoje naudojimo atvejų, norint gauti didesnės vertės duomenų rinkinius ar įžvalgas, reikia, kad šie nepriklausomi duomenų produktai tarpusavyje sąveikautų. Kad būtų galima įvykdyti norimas operacijas su duomenų produktais, duomenų tinklui įgyvendinti reikia valdymo modelį, kuris apima decentralizaciją ir domenų savarankiškumą, o svarbiausia, automatizuotą platformos sprendimų vykdymą. Tai ir vadinama konsoliduotu duomenų valdymu. Valdymo grupėms tenka sudėtingas darbas: išlaikyti pusiausvyrą tarp centralizacijos ir decentralizacijos. Kokie sprendimai turi būti lokalizuoti kiekvienoje srityje ir kokie sprendimai turėtų būti priimami visuotinai visoms sritims. Galiausiai visuotiniai sprendimai turi vieną tikslą - sukurti sąveikumą ir sudėtinį tinklo poveikį atrandant ir sudarant duomenų produktus. Duomenų tinklo valdymo prioritetai skiriasi nuo tradicinio analitinių duomenų valdymo. Nors abiejų valdymo tipų tikslas yra gauti duomenų vertę, tradicinis duomenų valdymas bando tai pasiekti centralizuojant sprendimų priėmimą ir nustatant visuotinį kanoninį duomenų atvaizdavimą su minimalia parama pokyčiams. Tuo tarpu duomenų tinklo duomenų valdymas apima pokyčius ir įvairius interpretavimo kontekstus.

4. Tyrimo metodologinė dalis

4.1 Tyrimo metodologija

Atliekant šį tyrimą buvo naudojami kelių tipų šaltiniai: pirminiai ir antriniai. Pirminis šaltinis – tyrimas, kuris buvo atliktas, siekiant išsiaiškinti problemas, su kuriomis susiduria organizacijos ir darbuotojai valdant ir tvarkant duomenis bei veiksnius, kurie imponuoja, kad organizacija galėtų implementuotis duomenų tinklo architektūrą ir jos poreikį. Antriniai duomenų šaltiniai buvo naudojami atliekant literatūros analizę, kurioje aprašytos duomenų saugyklos, duomenų ežero ir duomenų tinklo architektūros. Taip pat nagrinėjamos moksliniuose straipsniuose analizuojamos problemos, kurios iškyla valdant duomenis ir iššūkiai su kuriais susiduriama. Buvo pasirinktas kiekybinis tyrimo metodas, naudojant apklausos formą. Kiekybinis tyrimas yra paremtas skaičiumi žmonių, kurie atsako į apklausos klausimus. Tikslus žmonių skaičius ir jų išreikštos nuomonės tyrimą padaro labiau objektyvų ir lengviau apibendrinamą pagal išsikeltus kriterijus. Tai labai patogus būdas paneigti arba patvirtinti išsikeltas hipotezes, nes apklausoje pateikti klausimų blokai turi fiksuotus teiginius, kuriuos respondentai renkasi atsakydami (sutinka ar nesutinka) ir taip sudaro atskiras grupes atsakymų, kurie leidžia daryti tam tikras prielaidas ir koreliuoti tarp skirtingų veiksnių. (Qualitative versus Quantitative…)

Klausimynas buvo sudarytas naudojant „Google Forms“ įrankį. Klausimyną sudaro 10 klausimų (žr. 1 priede). Pirmasis klausimas padeda atfiltruoti atsakymus, kai respondentai pasirenka naudojamą technologiją. Antras, trečias ir ketvirtas klausimai – didieji blokai, kuriuos sudaro atitinkamai teiginių, kuriuos respondentai atsako pasirinkdami nuo 1 iki 5 skalėje (1 – visiškai nesutinku, 2 – nesutinku, 3 – nei sutinku, nei nesutinku, 4 – sutinku, 5 – visiškai sutinku). Penktas, šeštas, septintas, aštuntas, devintas ir dešimtas klausimai – demografiniai. Demografiniais klausimais buvo siekiama išgryninti apklaustuosius ir susidaryti aiškesnį vaizdą. Klausimyno struktūra pateikiama 1 lentelėje.

1 lentelė. **Klausimyno struktūra**

Klausimo tipas	**Klausimo sudėtis**	**Skalė**
Filtruojantis klausimas	Pasirenkama naudojama technologija (duomenų saugykla arba duomenų ežeras)	Plataus pasirinkimo skalė (šiuo atveju galima rinktis vieną iš dviejų arba abu)
Duomenų saugyklos blokas	Teiginiai apie duomenų saugyklos technologiją (problemos, ypatumai, iššūkiai)	Skalė nuo 1 iki 5 (1 – visiškai nesutinku, 5 – visiškai sutinku)

Duomenų ežero blokas	Teiginiai apie duomenų ežero architektūrą (problemos, ypatumai, iššūkiai)	Skalė nuo 1 iki 5 (1 – visiškai nesutinku, 5 – visiškai sutinku)
Duomenų valdymo iššūkių blokas	Teiginiai apie duomenų valdymo iššūkius ir organizacijos požiūrį į duomenų valdymą	Skalė nuo 1 iki 5 (1 – visiškai nesutinku, 5 – visiškai sutinku)
Demografiniai klausimai	1. Amžius 2. Išsilavinimas 3. Patirtis 4. Pareigybė 5. Organizacijos darbuotojų skaičius 6. Veiklos sritis	1. Plataus pasirinkimo klausimas 2. Plataus pasirinkimo klausimas 3. Laisvas tekstas 4. TAIP arba NE 5. Plataus pasirinkimo klausimas 6. Plataus pasirinkimo klausimas

Šaltinis: sudaryta autoriaus

Klausimynas pradedamas klausimu, kuris padeda atfiltruoti respondentus. Pirmajame klausime renkamasi tarp dviejų variantų: duomenų saugyklos ir duomenų ežero. Buvo pasirinkta duomenų saugyklos technologija, nes ji yra viena iš plačiausiai naudojamų duomenų saugojimo technologijų. 2021 metais, TDWI (angl. *Transforming Data With Intelligence*) atliktas tyrimas parodė, kad 53% organizacijų naudoja duomenų saugyklos technologiją fiziškai, o 36% debesyje. (Korolov, 2021). Tuo tarpu duomenų ežeras turbūt buvo viena iš labiausiai dėmesį prikausčiusių architektūrų 2010 metais, kai šį terminą pavartojo James‘as Dixon‘as (Foote, 2020). Atfiltravus respondentus ir jų naudojamą technologiją, galima žvelgti į kitus klausimų blokus, kurie parodo pagrindines problemas ir iššūkius su kuriais susiduria vienos ar kitos saugyklos naudotojai.

Antrasis klausimas – duomenų saugyklos blokas. Šis klausimas yra vienas iš trijų didžiųjų blokų. Jis susideda iš 15-os teiginių, kurie buvo parengti, remiantis Angel Ojeda-Castro, Mysore Ramaswamy, Angel Rivera-Collazo ir Ahmad Jumah parengtu straipsniu apie kritinius sėkmės faktorius implementuojant duomenų saugyklą (Castro, Ramaswamy, Collazo ir Jumah, 2011), Fernando Almeida ir Catalin Nicolae Calistru straipsniu apie pagrindinius iššūkius ir problemas valdant didžiuosius duomenis (Almeida ir Calistru, 2013) bei Umar Aftab ir Ghazanfar Farooq Siddiqui parengtu straipsniu apie didžiųjų duomenų valdymą naudojant duomenų saugyklą (Aftab ir Siddiqui, 2018). Klausime išvardinti teiginiai – problemos, su kuriomis dažniausiai susiduria darbuotojai ir iššūkiai su kuriais tenka tvarkytis. Šiuo klausimu siekta išsiaiškinti su kokiomis problemomis ir iššūkiais dažniausiai susiduria apklaustieji ir kaip kai kurie iš šių iššūkių galėtų būti sprendžiami naudojant duomenų tinklo architektūrą.

Trečiasis klausimas yra apie pagrindines problemas ir iššūkius su kuriais susiduriama naudojant duomenų ežero architektūrą. Sudarant teiginius buvo remiamasi Phint Phyu Khine ir Zhao Shun Wang parašytu straipsniu apie duomenų ežero ideologiją didžiųjų duomenų eroje (Khine ir Wang, 2018), Fatemeh Nargesian, Erkang Zhu, Renee J. Miller, Ken Q. Pu ir Patricia C. Arocena straipsniu apie duomenų ežero valdymą bei jo iššūkius ir galimybes (Nargesian, Zhu, Miller, Pu ir Arocena, 2019). Taip pat Carlos Ordonez, Gabriele Anderst-Kotsis, Ismail Khalil, Il-Yeol Song ir A Min Tjoa parašyta knyga – „Didžiųjų duomenų analitika ir žinių atradimas“ (Ordonez, Kotsis, Khalil, Song ir Tjoa, 2019). Šiuo klausimu siekta išsiaiškinti su kokiomis problemomis ir iššūkiais dažniausiai susiduria apklaustieji ir kaip kai kurie iš jų galėtų būti sprendžiami naudojant duomenų tinklo architektūrą.

Ketvirtasis klausimas yra apie problemas ir teiginius, susijusius su duomenų valdymu ir jų valdymo principais. Klausimą sudaro 22-u teiginiai, kuriais siekta išsiaiškinti su kokiomis bendrai problemomis susiduriama valdant duomenis ir su kokiais teiginiais sutinkama. Dalis teiginių buvo pateikti problemų principų, kai įvardijama problema ir respondentas renkasi ar su ja sutinka ar nesutinka. Kita dalis teiginių buvo pateikta apie, kokios duomenų valdymo procedūros yra taikomos organizacijoje, kokie duomenų valdymo metodai naudojami, kokie analitiniai įrankiai naudojami, kaip traktuojami duomenys ir t.t. Šiuo klausimu siekta susieti bendras problemas, su anksčiau paminėtomis problemomis, kai naudojama duomenų saugykla arba duomenų ežeras. Taip pat siekta įvertinti įmonės pasirengimą implementuotis duomenų tinklo architektūrą, kokie požymiai atitinka, ko trūksta ir iš kur atsiranda poreikis. Rengiant teiginius buvo remtasi literatūros analizėje naudotais šaltiniais apie duomenų valdymo procedūras, duomenų valdymo sistemas, duomenų kokybės standartus, duomenų analitikos įrankius bei duomenų tinklo architektūros principus. (Data Governance: An…)(Dehghani, 2019)

Galiausiai buvo parengti demografiniai klausimai. Jais buvo siekiama išsiaiškinti respondentų amžių, išsilavinimą, darbo patirtį organizacijoje, pareigybę, organizacijos darbuotojų skaičių bei organizacijos veiklos sritį. Iš šių šešių klausimų galima išskirti darbo patirtį, pareigybę bei organizacijos darbuotojų skaičių. Kuo didesnę darbo patirtį turintis ir kuo aukštesnes pareigas užimantis respondentas, tuo labai tikėtina, kad jo suteikta informacija ir atsakymai bus tikslesni, nei naujo organizacijos žmogaus ar žemesnes pareigas užimančio respondento. Taip pat labai svarbus punktas – organizacijos dydis. Jeigu organizacija yra maža, tikėtina, kad jai nebus poreikio implementuotis duomenų tinklo architektūrą. Suderinus visus klausimus su darbo vadovu ir gavus leidimą, klausimynas buvo išplatintas interneto pagalba.

4.2 Imties dydžio nustatymas ir duomenų apdorojimo metodas

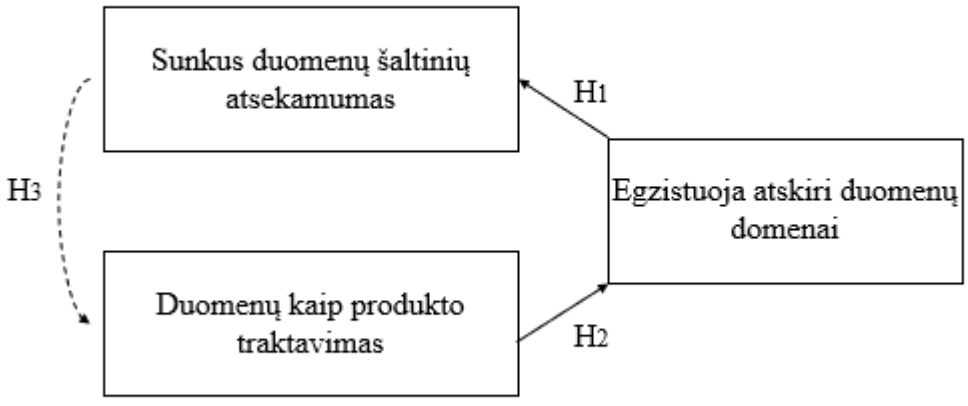
Atliekant tyrimą, duomenys buvo renkami pasirinkus tikslinės atrankos metodą. Toks metodas buvo pasirinktas dėl to, nes apklausos klausimai ir teiginiai buvo gana specifiniai ir reikalaujantys žinių toje srityje. Imties dydis buvo skaičiuotas naudojant (qualtrics.com, 2022) internetinę svetainę, kurioje naudojama statistinė imties dydžio radimo formulė. Tyrimo tikslinę auditoriją sudaro visi didelėse (>250 darbuotojų) organizacijose dirbantys asmenys, kurie yra susidūrę/prislietę arba turintys praktinių žinių apie duomenų saugyklą arba duomenų ežerą. Taip pat pageidautina, kad asmuo užimtų vadovaujančias pareigas. Naudojantis Viešo įmonių registro duomeninis ir atsifiltravus įmones pagal darbuotojų skaičių ir industrines sritis, kurios labiausiai tikėtina bus aktualios (transporto, informacijos ir ryšių/IT, finansinė ir draudimo veikla, švietimas ir mokslas) buvo gautos 153 kriterijus atitinkančios įmonės. Kadangi tikslaus skaičiaus, kiek žmonių sudaro IT departamentus arba kiek tiksliai žmonių dirba būtent prie IT veiklos ar duomenų saugyklų/duomenų ežero, nustatyti praktiškai neįmanoma, buvo pasirinktas variantas į imtį įtraukti visus įmonėje esančius darbuotojus. Susumavus bendrą darbuotojų skaičių buvo gauta 120 064 asmenų suma. Šis skaičius buvo naudojamas nustatant imties dydį, kai patikimumo lygis – 95% ir paklaidos procentas – 8%. Suvedus skaičius į formulę buvo gautas imties dydis – 150.

Tyrimo rezultatams rezultatams apdoroti ir analizuoti buvo naudojami du įrankiai: Microsoft Excel ir IBM SPSS Modeler. Visų pirma Microsoft Excel pagalba (naudojantis „IF“ funkcija) buvo užkoduoti iš apklausos gauti respondentų atsakymai. Kadangi respondentai į pagrindinius klausimų blokus, kurie buvo paremti teiginiais iš literatūros šaltinių, atsakinėjo skalėje nuo 1 iki 5, buvo nuspręsta šiuos pasirinkimus (1-2-3-4-5) užkoduoti. Atsakymai, kurie buvo pasirinkti nuo 1 iki 3 buvo koduojami kaip: „Ne“. Atsakymai, kurie buvo pasirinkti nuo 4 iki 5 buvo koduojami kaip: „Taip“. Toks kodavimo principas buvo pasirinktas todėl, nes analizuojant veiksnius, pagal pasirinktą tyrimo metodą, kurie lemia duomenų tinklo architektūros poreikį ir nusako duomenų valdymo iššūkius buvo patogiau analizuoti turint du kintamuosius (Taip arba Ne). Respondentai buvo užkoduoti kaip „Respondento ID“ (1,2,3,4,5…), naudojamas technologijos tipas (duomenų saugykla arba duomenų ežeras) koduojami kaip „DS“ ir „DE“ (atitinkamai duomenų saugykla ir duomenų ežeras). Demografiniai duomenys buvo palikti tokie, kokie gauti iš apklausos. IBM SPSS Modeler įrankio pagalba buvo naudojami du metodai: dvi asociacijų taisyklės (Apriori ir Carma). Apriori funkcijos pagalba buvo ieškomos asociacijos tarp skirtingų duomenų rinkinių ir matuojamos jų palaikymo (angl. support%) ir patikimumo (angl. confidence%) procentinės išraiškos. Tokiu būdu siekta iššiaiškinti skirtingus ir įdomius ryšius tarp skirtingų veiksnių rinkinių bei jų sąveikavimą tarpusavyje. Naudojant funkciją sudaromos taisyklės,

kurios sudaromos pagal pasirinktas reikšmes „consequent“ ir „antecedent“ stulpeliuose. Šiame kontekste reiškia, kad „consequent“ yra „then“ taisyklės dalis, o „antecedent“ yra „if“ taisyklės dalis. Pavyzdžiui: „**If** (antecedent stulpelis) <pasirinktas veiksnys>, **then** (consequent stulpelis) <pasirinktas veiksnys>“. Tada yra skaičiuojamas respondentų skaičius procentais (support%), kurie pasirinko „antecedent“ stulpelyje esantį veiksnį ir patikimumo procentas (confidence%), kuris nurodo koks procentas atitinka būtent tokį pasirinkimą. Palaikymo procentinė išraiška nurodo respondentų skaičių procentais, kurie pasirinko „antecedent“ stulpelyje esantį veiksnį/veiksnius, o patikimumo procentinė išraiška nurodo veiksnių derinio tikslumą, atsižvelgiant į „consequent“ stulpelio pasirinktas reikšmes. Apriori funkcijos privalumas yra tai, kad kairėje arba dešinėje stulpelio pusėje gali būti atvaizduojamas tik 1 veiksnys, o kitoje >1 ir atvirkščiai. Taip pat buvo naudojama Carma funkcija, kuri yra labai panaši į Apriori, turi tuos pačius lyginimo veiksnius (palaikymas ir patikimumas), tačiau skiriasi tuom, kad kairės ir dešinės pusės stulpeliai nėra apriboti skaičiumi. Jeigu yra poreikis ir apibrėžiamos tokios taisyklės, tiek vienoje, tiek ir kitoje pusėje gali būti >1 veiksnys. Ši funkcija padėjo atsifiltruoti skirtingus veiksnius ir pamatuoti jų sąveikavimą tarpusavyje su kitais veiksniais.

Atliekant tyrimą buvo užtikrintas visiškas respondentų anonimiškumas. Respondentai buvo supažindinti su apklausos tikslu, tema ir duomenų panaudojimo tikslu.

4.3 Hipotezių vystymas



6 pav. Tyrimo koncepcija (šaltinis: sudaryta autoriaus)

1. *Duomenų šaltinių atsekamumas*

Duomenų šaltinių atsekamumo problema yra viena iš dažnai pasitaikančių duomenų valdymo problemų. Tinkamas duomenų šaltinių atsekamumas gali padėti užtikrinti tinkamą duomenų kokybę, jų patikimumą bei aiškų duomenų kelią. Egzistuojant tokiam duomenų kiekiui, jų neįmanoma saugoti vienoje vietoje ar kokiame nors kietąjame diske. Įprastai duomenys būna pasiskirstę po įvairias repozitorijas, pradedant fizinėmis saugyklomis ir baigiant duomenimis nutolusiais nuo fizinės savininko vietos. (L'Heureux, Grolinger, Elyamany ir Capretz, 2017) Duomenų šaltinių atsekamumo problemą galimai padėtų išspręsti domenais paremta architektūra. Komanda, kuri turėtų nuosavybės teisę (angl. data ownership) į tam tikrą duomenų domeną ir būtų atsakingą už jo priežiūrą, galėtų fokusuotis tik į to domeno duomenų šaltinių atsekamumą.

H_1: Atskirų duomenų domenų nebuvimas teigiamai veikia duomenų šaltinių atsekamumo problemą.

2. *Duomenys kaip produktas*

Vienas iš esamų analitinių duomenų architektūrų iššūkių - didelės išlaidos, susijusios su kokybiškų duomenų atradimu, supratimu, patikimumu ir galiausiai naudojimu. Jei ši problema nesprendžiama, didėjant duomenų kiekiui ji tik aštrėja, nes daugėja vietų ir komandų, teikiančių duomenis. Duomenų kaip produkto principas skirtas duomenų kokybės ir senos duomenų siloso problemos sprendimui. Analitiniai duomenys, kuriuos teikia sritys, turi būti traktuojami kaip produktas, o šių duomenų naudotojai turi būti traktuojami kaip klientai. (Dehghani, 2020) Tačiau šia hipoteze siekta ne tik paneigti arba patvirtinti faktą, kad duomenų kaip produkto traktavimas įgalina domenais paremtą architektūrą, bet ir paanalizuoti skirtingus duomenų rinkinius, kurie galbūt įrodytų, kad net ir duomenų traktavimas, kaip produktas nėra lemiamas veiksnys domenų architektūrai.

H_2: Duomenų kaip produkto traktavimas teigiamai veikia atskirų duomenų domenų egzistavimą.

3. *Sustiprinančioji hipotezė H_3 (Sunkus duomenų šaltinių atsekamumas → Duomenys traktuojami kaip produktas)*

Ši hipotezė nėra paremta literatūros šaltiniais, tačiau atsižvelgiant į tyrimo klausimus ir veiksnius, daroma prielaida, kad ji gali padėti sustiprinti paneigtą arba pasitvirtinusią hipotezę. Jeigu H_1 ir H_2

pasitvirtintų arba būtų paneigtos, būtų galima pamėginti išgauti tokią hipotezės(-ių) baigtį nulėmusių sustiprinančiųjų įžvalgų.

5. Tiriamoji tyrimo dalis

5.1 Demografinių duomenų apžvalga

2 lentelė. **Demografinių klausimų atsakymų pasiskirstymas**

Kintamasis	Klausimo tipas	Respondentų sk. %
Amžius	18-25m.	32,81 %
	26-35m.	46,88 %
	36-45m.	17,19 %
	46-55m.	3,21 %
	56 ir daugiau	0 %
Išsilavinimas	Pradinis	0 %
	Pagrindinis (nebaigtas vidurinis)	0 %
	Vidurinis	9,38 %
	Profesinis (profesinė mokykla, vidurinis su profesine kvalifikacija)	0 %
	Aukštesnysis (technikumas, aukštesniosios mokyklos)	6,25 %
	Neuniversitetinis aukštasis (kolegija)	0 %
	Universitetinis aukštasis (bakalauro laipsnis)	34,38 %
	Universitetinis aukštasis (magistro laipsnis, ar jam prilygstanti profesinė kvalifikacija/mokslų daktaras)	50 %
Pareigos (ar pareigos vadovaujančios?)	Taip	61 %
	Ne	39 %
Darbuotojų skaičius organizacijoje	Mažiau už 250	20 %
	Daugiau arba lygu 251	80 %
Veiklos sritis	Žemės ūkis, žuvininkystė	0 %
	Energetika	0 %
	Statyba	0 %
	Gamyba	6,25 %
	Didmeninė ir mažmeninė prekyba	7,81 %
	Transportas	4,69 %
	Apgyvendinimo ir maitinimo paslaugos	0 %
	Informacija ir ryšiai, IT	50 %
	Finansinė ir draudimo veikla	14,1 %
	Paslaugų verslas	3,13 %
	Švietimas, mokslas	7,8 %

	Sveikatos priežiūra	6,25 %
	Kultūra	0 %

Šaltinis: sudaryta autoriaus

5.2 Hipotezių testavimas

3 lentelė**. H1: Atskirų duomenų domenų nebuvimas teigiamai veikia duomenų šaltinių atsekamumo problemą**.

Consequent	Antecedent	Support (%)	Confidence (%)
Sunkus duomenų šaltinių atsekamumas „Taip“	Egzistuoja atskiri duomenų domenai „Taip“	23,44%	86,67%
Sunkus duomenų šaltinių atsekamumas „Taip“	Egzistuoja atskiri duomenų domenai „Ne“	76,56%	67,35%
Sunkus duomenų šaltinių atsekamumas „Ne“	Egzistuoja atskiri duomenų domenai „Ne“	76,56%	32,65%
Sunkus duomenų šaltinių atsekamumas „Ne“	Egzistuoja atskiri duomenų domenai „Taip“	23,44%	13,33%

Šaltinis: sudaryta autoriaus

Testuojant H1 ir H2 hipotezes buvo naudojamas Apriori metodas. Apriori metodo pagalba buvo išfiltruoti du veiksniai: „sunkus duomenų šaltinių atsekamumas“ ir „egzistuoja atskiri duomenų domenai“. Metodo modelio nustatymuose buvo nustatyta minimali palaikymo reikšmė 5% ir minimali patikimumo reikšmė 5%, kad būtų išrinkti visi įmanomi variantai. Gauti veiksnių pasirinkimo duomenys pavaizduoti lentelėje (žr. 3 lentelė). Priminimui, prie veiksnių esantys indikatoriai „Taip“ arba „Ne“, reiškia respondentų kodavimą pagal jų pasirinkimus apklausos skalėje nuo 1 iki 5. 1,2 arba 3 pasirinkę respondentai užkoduoti kaip „Ne“, 4 arba 5 kaip „Taip“. (žr. 4.2 „Imties dydžio nustatymas ir duomenų apdorojimo metodas“)

Remiantis lentelės duomenimis, buvo nustatyta, kad mažiausiai pasitvirtinanti veiksnių pora yra „Sunkus duomenų šaltinių atsekamumas (Ne) “ ir „Egzistuoja atskiri duomenų domenai (Taip)“. Iš visų 64 atsakiusių respondentų 23,44% rinkosi, kad egzistuoja atskiri duomenų domenai, tačiau šio derinio patikimumo procentas mažiausias iš visų 4 derinių. Egzistuoja tik 13,33% patikimumo rodiklis, kad respondentai pasirinkę, kad egzistuoja atskiri duomenų domenai rinkosi, kad nėra duomenų šaltinių atsekamumo problemos. Kitas veiksnių derinys „Sunkus duomenų šaltinių atsekamumas (Ne)“ ir „Egzistuoja atskiri duomenų domenai (Ne)“. Šio derinio palaikymo procentas 76,56%, o patikimumo rodiklis 32,65%. Gana didelė dalis respondentų (76,56%) rinkosi, kad jų organizacijoje neegzistuoja

atskiri duomenų domenai. Tačiau įdomu tai, kad patikimumo procentas nėra didelis, kad neegzistuojant atskiriems duomenų domenams, neegizstuoja ir duomenų šaltinių atsekamumo problema. Todėl atsižvelgiant į patikimumo ir palaikymo procentinį santykį, galima preziumuoti, kad neegzistuojant atskiriems duomenų domenams, turėtų būti susiduriama su duomenų šaltinių atsekamumo problema. Labiausiai subalansuota veiksnių pora yra „Sunkus duomenų šaltinių atsekamumas (Taip)“ ir „Egzistuoja atskiri duomenų domenai (Ne)“. Palaikymas (76,56%) ir patikimumas (67,35%). Nors patikimumo procentas nėra pats aukščiausias rezultatų lentelėje, tačiau balansas tarp palaikymo rodiklio ir patikimumo rodiklio yra labiausiai išreikštas. Remiantis duomenimis, galima teigti, kad didžioji dauguma respondentų susiduria su duomenų šaltinių atsekamumo problema, kai organizacijoje neegzistuoja atskiri duomenų domenai. Ir paskutinis neaptartas veiksnių derinys – „Sunkus duomenų šaltinių atsekamumas (Taip)“ ir „Egzistuoja atskiri duomenų domenai (Taip)“. Šio derinio rodikliai: palaikymas – 23,44%, o patikimumas – 86,67%. Nors palaikymas ir nėra didelis, patikimumo rodiklis yra aukščiausias iš lentelės duomenų. Toks rezultatas imponuoja, kad nors organizacijoje ir egzistuoja atskiri duomenų domenai, duomenų šaltinių atsekamumo problemos rizika išlieka. Remiantis tokiu rezultatu, galima daryti prielaidą, kad skirtingi duomenų domenai ne visada išsprendžia duomenų šaltinių atsekamumo problemą. Tačiau, kad tai patvirtinti, reikalingas didesnis respondentų atsakymų skaičius ir koncentruota tikslinė auditorija.

Atsižvelgus į rezultatus, galima teigti, kad atskirų duomenų domenų nebuvimas teigiamai veikia duomenų šaltinių atsekamumo problemą, nes gautas palaikymo ir patikimumo santykis buvo labiausiai subalansuotas tarp apklaustųjų.

4 lentelė. **H2: Duomenų kaip produkto traktavimas teigiamai veikia atskirų duomenų domenų egzistavimą.**

Consequent	Antecedent	Support (%)	Confidence (%)
Duomenys organizacijoje traktuojami kaip produktas „Ne“	Egzistuoja atskiri duomenų domenai „Ne“	76,56%	93,88%
Duomenys organizacijoje traktuojami kaip produktas „Taip“	Egzistuoja atskiri duomenų domenai „Taip“	23,44%	80%
Duomenys organizacijoje traktuojami kaip produktas „Ne“	Egzistuoja atskiri duomenų domenai „Taip“	23,44%	20%
Duomenys organizacijoje traktuojami kaip produktas „Taip“	Egzistuoja atskiri duomenų domenai „Ne“	76,56%	6,12%

Šaltinis: sukurta autoriaus

Norint patikrinti H2, kad duomenų kaip produkto traktavimai teigiamai veikia atskirų duomenų domenų egzsitavimą, taip pat buvo naudojamas Apriori metodas. Metodo modelio nustatymuose buvo nustatyta minimali palaikymo reikšmė 5% ir minimali patikimumo reikšmė 5%, kad būtų išrinkti visi įmanomi variantai. Gauti veiksnių pasirinkimo duomenys pavaizduoti lentelėje (žr. 4 lentelė).

Remiantis lentelės duomenimis, matoma, kad mažiausias balansas užfiksuotas tarp šių veiksnių: „Duomenys organizacijoje traktuojami kaip produktas (Taip)“ ir „Egzistuoja atskiri duomenų domenai (Ne)“. Šio derinio rodikliai: palaikymas (76,56%), o patikimumas viso labo 6,12%. Atsižvelgiant į rezultatą, matosi, kad neegzistuojant atskiriems duomenų domenams yra maža tikimybė, kad duomenys organizacijoje traktuojami kaip produktas. Kita veiksnių pora: „Egzistuoja atskiri duomenų domenai (Taip)“ ir „Duomenys organizacijoje traktuojami kaip produktas (Ne)“. Palaikymo rodiklis – 23,44%, o patikimumo rodiklis – 20%. Gana mažas respondentų procentas rinkosi, kad organizacijoje egzistuoja atskiri duomenų domenai, o to pasekoje, mažai tikėtina, kad duomenys organizacijoje nėra traktuojami kaip produktas. Didelis patikimumo procentas (80%) užfiksuotas tarp veiksnių: „Egzistuoja atskiri duomenų domenai (Taip)“ ir „Duomenys traktuojami kaip produktas (Taip)“. Palaikymo procentas – 23,44%. Tai reiškia, kad nors ir nedidelė dalis respondentų pažymėjo, kad organizacijoje egzistuoja atskiri duomenų domenai, bet jeigu jie egzistuoja, tai didelė tikimybė, kad duomenys organizacijoje traktuojami kaip produktas. Didžiausi procentiniai rodikliai užfiksuoti tarp veiksnių: „Egzistuoja atskiri duomenų domenai (Ne)“ ir „Duomenys organizacijoje traktuojami kaip produktas (Ne)“. Palaikymo rodiklis – 76,56%, o patikimumo rodiklis net 93,88%. Tai reiškia, kad jeigu organizacijoje neegzistuoja atskiri duomenų domenai, tai duomenys nėra traktuojami kaip produktas. Šis veiksnių derinys yra lyg atvirkštinis H2 patvirtinimas, nes jeigu neegzistuojant atskiriems duomenų domenams duomenys nėra traktuojami kaip produktas, tikėtina, kad egzistuojant atskiriems duomenų domenams, duomenys būtų traktuojami kaip produktas.

Apžvelgus rezultatus, galima teigti, kad duomenų kaip produkto traktavimas teigiamai veikia atskirų duomenų domenų egzistavimą.

5 lentelė. ***H3: Sustiprinančioji hipotezė H3 (Sunkus duomenų šaltinių atsekamumas → Duomenys traktuojami kaip produktas)***

Consequent	Antecedent	Support (%)	Confidence (%)
Sunkus duomenų šaltinių atsekamumas „Taip“	Duomenys organizacijoje traktuojami kaip produktas „Taip“	23,44%	80%
Sunkus duomenų šaltinių atsekamumas „Taip“	Duomenys organizacijoje traktuojami kaip produktas „Ne“	76,56%	69,39%
Sunkus duomenų šaltinių atsekamumas „Ne“	Duomenys organizacijoje traktuojami kaip produktas „Ne“	76,56%	30,61%
Sunkus duomenų šaltinių atsekamumas „Ne“	Duomenys organizacijoje traktuojami kaip produktas „Taip“	23,44%	20%

Šaltinis: sukurta autoriaus

Pasitvirtinus H_1 ir H_2 galima apžvelgti sustiprinančiosios hipotezės lentelės duomenis. Skaičiuojant lentelės rezultatus buvo naudojamas Apriori metodas. Metodo modelio nustatymuose buvo nustatyta minimali palaikymo reikšmė 5% ir minimali patikimumo reikšmė 5%, kad būtų išrinkti visi įmanomi variantai. Gauti veiksnių pasirinkimo duomenys pavaizduoti lentelėje (žr. 5 lentelė).

Remiantis lentelės duomenimis, matoma, kad didžiausias balansas tarp palaikymo ir patikimumo reikšmių užfiksuotas tarp šių veiksnių: „Duomenys organizacijoje traktuojami kaip produktas (Ne)“ ir „Sunkus duomenų šaltinių atsekamumas (Taip)“. Šio veiksnių derinio rodikliai: palaikymas – 76,56%, o patikimumo rodiklis – 69,39%. Rezultatai nurodo, kad kai duomenys nėra traktuojami kaip produktas, susidaro didelė tikimybė, kad egzistuoja duomenų šaltinių atsekamumo problema. Lyginant lentelių (3,4,5) duomenis, matoma bendra tendencija. 3-oje lentelėje labiausiai subalansuotas veiksnių derinys buvo: „Egzistuoja atskiri duomenų domenai (Ne)“ ir „Sunkus duomenų šaltinių atsekamumas (Taip)“ (76,56% palaikymas ir 67,35% patikimumas). 4-oje lentelėje labiausiai subalansuotas veiksnių derinys buvo: „Duomenys organizacijoje traktuojami kaip produktas (Ne)“ ir „Egzistuoja atskiri duomenų domenai (Ne)“ (76,56% palaikymas ir 93,88% patikimumas). Apibendrinus šias tris lenteles matosi bendras požymis, kad kai duomenys organizacijoje nėra traktuojami kaip produktas ir neegizstuoja atskiri duomenų domenai, tai atsiranda ir duomenų šaltinių atsekamumo problema.

Galiausiai, apžvelgus rezultatus, galima teigti, kad sustiprinančioji hipotezė pasitvirtino ir pasitvirtinusios hipotezės H_1 ir H_2 tapo stipriau pagrindžiamos.

6 lentelė. **Hipotezių apžvalga**

Hipotezė:	Rezultatas:
H1: Atskirų duomenų domenų nebuvimas teigiamai veikia duomenų šaltinių atsekamumo problemą.	Pasitvirtino
H2: Duomenų kaip produkto traktavimas teigiamai veikia atskirų duomenų domenų egzistavimą.	Pasitvirtino
Sustiprinančioji hipotezė H3 (Sunkus duomenų šaltinių atsekamumas → Duomenys traktuojami kaip produktas)	Pasitvirtino

Šaltinis: sudaryta autoriaus

5.3 Tyrimo duomenų interpretacija

7 lentelė. **Dažniausiai pasikartojantys veiksnių rinkiniai pagal aukščiausius palaikymo ir patikimumo rodiklius**

Consequent	Antecedent	Support (%)	Confidence (%)
Technologija: Duomenų saugykla	Duomenys organizacijoje traktuojami kaip produktas „Ne“ Egzistuoja atskiri duomenų domenai „Ne“	71,88%	100%
Sunkus duomenų savininkų atsekamumas „Taip“	Sunkus duomenų šaltinių atsekamumas „Taip“ Technologija: Duomenų saugykla	65,63%	100%
Sunkus duomenų šaltinių atsekamumas „Taip“ Veiklos procesų automatizavimo trūkumas „Taip“	Negarantuotas sistemos našumas „Taip“	66,13%	90,24%
Duomenų darnos (integracijų) trūkumas „Taip“	Sudėtingas duomenų šaltinių atsekamumas „Taip“	75,80%	82,98%
Negarantuotas sistemos našumas „Taip“ Sudėtinga sudaryti tikslius įmonės pajėgumo planus, siekiant konkrečių tikslų „Taip“	Automatizacijos trūkumas (rankiniai sistemos ir duomenų atnaujinimai) „Taip“ Sudėtingas duomenų šaltinių atsekamumas „Taip“	66,13%	87,80%

Šaltinis: sudaryta autoriaus

Siekiant įvairiapusiškiau paanalizuoti tyrimo metu gautus rezultatus, buvo panaudotas Carma metodas. Metodo modelio nustatymuose buvo nustatyta minimali palaikymo reikšmė 65% ir minimali patikimumo reikšmė 80%, kad būtų išrinkti visi dažniausiai pasikartojanys ir stipriausius rodiklius fiksuojantys variantai. Gautų rezultatų duomenys pavaizduoti lentelėje (žr. 7 lentelė).

Buvo išskirti 5 veiksnių variantai, kurie fiksavo aukščiausius rodiklius ir balansą tarpusavyje.

1. „Duomenys organizacijoje traktuojami kaip produktas (Ne) + Egzistuoja atskiri duomenų domenai (Ne)“ ir „Technologija: duomenų saugykla“. Šio veiksnių derinio rodikliai: palaikymas – 71,88%, patikimumas – 100%. Remiantis rezultatais, buvo gauta, kad kai duomenys organizacijoje nėra traktuojami kaip produktas ir neegzistuoja atskiri duomenų domenai – naudojama duomenų saugyklos technologija.

2. „Sunkus duomenų šaltinių atsekamumas (Taip) + Technologija: duomenų saugykla“ ir „Sunkus duomenų savininkų atsekamumas (Taip)“. Šio veiksnių derinio rodikliai: palaikymas – 65,63%, patikimumas – 100%. Rezultatai rodo, kad naudojant duomenų saugyklos technologiją ir esant duomenų šaltinių atsekamumo problemai, egzistuoja ir duomenų savininkų atsekamumo problema.

3. „Negarantuotas sistemos našumas (Taip)“ ir „Sunkus duomenų šaltinių atsekamumas (Taip) + Veiklos procesų automatizavimo trūkumas (Taip)“. Šio veiksnių derinio rodikliai: palaikymas – 66,13%, patikimumas – 90,24%. Rezultatai rodo, kad esant negarantuotam sistemos našumui egzistuoja duomenų šaltinių atsekamumo problema bei veiklos automatizavimo trūkumo problema.

4. „Sudėtingas duomenų šaltinių atsekamumas (Taip)“ ir „Duomenų darnos (integracijų) trūkumas (Taip)“. Šio veiksnių derinio rodikliai: palaikymas – 75,80%, patikimumas – 82,98%. Rodikliai nurodo, kad esant duomenų šaltinių atsekamumo problemai, tikėtinas duomenų darnos trūkumas.

5. „Automatizacijos trūkumas (rankiniai sistemos ir duomenų atnaujinimai) (Taip) + Sudėtingas duomenų šaltinių atsekamumas (Taip)“ ir „Negarantuotas sistemos našumas (Taip) + Sudėtinga sudaryti tikslius įmonės pajėgumo planus, siekiant konkrečių tikslų (Taip)“. Šio veiksnių derinio rodikliai: palaikymas – 66,13%, patikimumas – 87,80%. Rodiklių rezultatas rodo, kad egzistuojant automatizacijos trūkumo ir duomenų šaltinių atsekamumo problemai, susiduriama su negarantuotu sistemos našumu ir iššūkiu sudaryti tikslius įmonės pajėgumo planus.

Išvados ir rekomendacijos

1. Apibendrinus tyrimo rezultatus, matosi tendencija, kad didžioji dauguma duomenų valdymo problemų būdinga duomenų saugyklai. Duomenų savininkų ir šaltinių atsekamumas, integracijų trūkumas, sistemos našumo problemos, tikslių organizacijos pajėgumo planų sudarymo problema ir t.t., kyla naudojant duomenų saugyklos technologiją. Tai leidžia manyti, kad nors duomenų saugyklos technologija yra plačiai naudojama ir yra pakankamai gerai išvystyta, didėjant duomenų kiekiui jos efektyvumas valdant ir analizuojant duomenis mažėja.
2. Tyrimo rezultatai parodė, kad rinkoje jau atsiranda organizacijų, kurios duomenis traktuoja kaip produktą ir egzistuoja atskiri duomenų domenai. Šie veiksniai yra vieni pagrindinių, siekiant implementuoti duomenų tinklo architektūrą. Tai leidžia manyti, kad įmonės nors ir ne sparčiai, bet mažais žingsniais eina link to, kad nuo tradicinių centralizuotų duomenų saugojimo repozitorijų pereitų prie decentralizuotų repozitorijų ir/arba atsirastų duomenų sričių savininkai.
3. Remiantis literatūros analizės pagalba gautomis įžvalgomis, galima spręsti, kad be efektyvaus duomenų valdymo organizacijai praktiškai neįmanoma funkcionuoti. Vis plačiau naudojamos duomenų ir informacijos valdymo architektūros, kurios leidžia tiksliai apsibrėžti taisykles ir procedūras. Duomenų valdymo svarbą patvirtino ir tyrimo metu surinkti atsakymai, kurie teigė, kad esant duomenų savininkų ar šaltinių atsekamumo problemai, nepakankamam sistemos resursų kiekiui duomenims apdoroti ar duomenų darnos trūkumui – susiduriama su iššūkiais, kurie lemia apsunkintą planavimą, lemia sistemos našumo trūkumą, duomenų kokybės lygį ir kitus veiksnius.
4. Žvelgiant į praeities tendencijas, matoma, kad duomenų kiekis nuolatos auga, o juos reikia tinkamai apdoroti, saugoti ir interpetuoti. Atlikus tyrimą, galima pateikti rekomendaciją organizacijoms, kurios dirba su dideliais duomenų srautais. Kadangi tyrimo rezultatai parodė, kad atskirų duomenų domenų egzsistavimas eliminuoja ne vieną problemą, pasiūlymas būtų pradėti eiti ta linkme, kad centralizuotą duomenų valdymą pakeisti į decentralizuotą. Egzistuojant atskiriems domenams sumažėja jau nuo seno egzistuojanti duomenų siloso problema, kai reikalingi duomenys būna neprieinami arba valdomi duomenų savininko, kuris neturi pakankamai žinių, kad juos tinkamai interpretuoti. Taip pat atsiranda aiškiai išskirtos rolės, kokios organizacijos funkcijos ir komandos yra atsakingos už tam tikrus procesus. Ir galiausiai egzistuojant atskiriems domenams atsiranda galimybė duomenis traktuoti kaip produktą, suteikiant duomenų savininkui duomenų nuosavybės teisę. Tai pasitarnauja norint išspręsti

duomenų savininkų ir šaltinių atsekamumo problemas, nes komandos ar pavieniai asmenys, kurie valdo tam tikros srities duomenis, koncentruojasi tik į savo valdomų duomenų analizę ir apdorojimą.

5. Pasiūlymas ateities tyrimams: atliekant tokio tipo tyrimą, būtų geriau naudoti labiau koncentruotą tikslinę rinką ir surinkti kuo daugiau ekspertinių nuomonių. Kadangi apklausa yra gana specifinė, o klausimai sudarantys klausimyną reikalauja teorinių arba praktinių žinių tiek apie duomenų saugojimo technologijas, tiek apie patį duomenų valdymą, atsiranda iššūkis surinkti reikiamą skaičių atsakymų, kad padaryti tyrimą reikšmingu. Tokio tipo tyrimui taip pat labai pasitarnautų interviu metu gautos ekspertų įžvalgos ir gal net padėtų parengti kokybiškesnį ir tikslesnį klausimyną.

Literatūros sąrašas

1. Aftab, U. ir Siddiqui, G.F. (2018) Big Data Augmentation with Data Warehouse: A Survey. Prieiga per internetąhttps://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8622206
2. Almeida, F. ir Calistru, C.N. (2013) The main challenges and issues of big data management. Prieiga per internetą https://www.researchgate.net/profile/Fernando-Almeida-10/publication/272696610_The_main_challenges_and_issues_of_big_data_management/links/570251e108ae9969f70271c6/The-main-challenges-and-issues-of-big-data-management.pdf
3. Banafa, A. (2021) What is Data Lake? Prieiga per internetą https://www.bbvaopenmind.com/en/technology/digital-world/data-lake-an-opportunity-or-a-dream-for-big-data/
4. Bekker, A. (2019) Your Complete Guide to a Cloud Data Warehouse. Prieiga per internetą https://www.scnsoft.com/blog/cloud-data-warehouse
5. Belghith, O., Skhiri, S., Zitoun, S. ir Ferjaoui, S. (2021) A Survey of Maturity Models in Data Management. Prieiga per internetą https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9476197
6. Calzon, B. (2022). Discover Data Warehouse & Business Intelligence Architecture. https://www.datapine.com/blog/data-warehousing-and-business-intelligence-architecture/
7. Castro, A.O., Ramaswamy, M., Collazo, A.R. ir Jumah, A. (2011) CRITICAL FACTORS FOR SUCCESSFUL IMPLEMENTATION OF DATA WAREHOUSE. Prieiga per internetą https://iacis.org/iis/2011/88-96_AL2011_1611.pdf
8. Data Governance: An Essential Guide – 5 principles, 10 components, and getting started. Prieiga per internetą https://satoricyber.com/data-governance/essential-guide/
9. Data Management vs. Data Governance: The Difference Explained. Prieiga per internetą https://www.tableau.com/learn/articles/data-management-vs-data-governance
10. Data Warehouse Architecture. Prieiga per internetą https://www.javatpoint.com/data-warehouse-architecture
11. Data Warehouse Architecture: Traditional vs. Cloud. Prieiga per internetą https://panoply.io/data-warehouse-guide/data-warehouse-architecture-traditional-vs-cloud/
12. Data Warehouse vs. Data mart: a comparison. Prieiga per internetą https://www.stitchdata.com/resources/data-warehouse-data-mart
13. Data Warehouse: Practical Advice from the Experts. Prieiga per internetą https://pdfs.semanticscholar.org/81a2/a1299b70ce55c484f765ebc59c2dc24adab8.pdf

14. Data Warehousing Tutorials. (1-2p., 25-30p.) Prieiga per internetą http://index-of.es/Varios-2/Data%20Warehousing%20Tutorial.pdf
15. Dehghani, Z. (2019) How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh. Prieiga per internetą https://martinfowler.com/articles/data-monolith-to-mesh.html
16. Dehghani, Z. (2020) Data Mesh Principles and Logical Architecture. Prieiga per internetą https://martinfowler.com/articles/data-mesh-principles.html
17. Difference between Operational Database and Data Warehouse. Prieiga per internetą https://www.javatpoint.com/operational-database-vs-data-warehouse
18. Director of Emerging Technologies. Prieiga per internetą https://www.thoughtworks.com/profiles/z/zhamak-dehghani
19. Fan, W. ir Geerts, F. (2012) Foundations of Data Quality Management. Prieiga per internetą https://www.morganclaypool.com/doi/abs/10.2200/S00439ED1V01Y201207DTM030
20. Foote, K.D. (2018) A Brief History of the Data Warehouse. Prieiga per internetą https://www.dataversity.net/brief-history-data-warehouse/
21. Ford, N., Parsons, R. ir Kua, P. Building Evolutionary Architectures. Prieiga per internetą https://www.oreilly.com/library/view/building-evolutionary-architectures/9781491986356/ch04.html
22. Gartner Glossary. Dark Data. Prieiga per internetą https://www.gartner.com/en/information-technology/glossary/dark-data
23. Griffith, E. (2022) What Is Cloud Computing? Prieiga per internetą https://www.pcmag.com/news/what-is-cloud-computing
24. Grillenberg, A. ir Romeike, R. Teaching Data Management: Key Competencies and Opportunities. Prieiga per internetą https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/8264/file/cid07_S133-150.pdf
25. Henderson, D., Earley, S., Coleman, L.S., Sykora, E. ir Smith, E. (2017) Data Management Book of Knowledge 2nd Edition.
26. Information Architecture Basics. Prieiga per internetą https://www.usability.gov/what-and-why/information-architecture.html
27. Karatas, G. (2022) Data Lake: What it is, What are the Benefits & Challenges. Prieiga per internetą https://research.aimultiple.com/data-lake/
28. Khine, P.P. ir Wang, Z.S. (2018) Data Lake: a new ideology in big data era. Prieiga per internetą https://www.itm-conferences.org/articles/itmconf/pdf/2018/02/itmconf_wcsn2018_03025.pdf

29. King, T. (2021) The 14 Best Metadata Management Tools for 2021. Prieiga per internetą https://solutionsreview.com/data-management/the-best-metadata-management-tools/
30. Kiran, M., Murphy, P., Monga, I., Dugan, J. Ir Baveja, S.S. (2015) Lambda Architecture for Cost-effective Batch and Speed Big Data processing. Prieiga per internetą https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7364082
31. Korolov, M. (2021) Top 5 elements needed for a successful data warehouse. Prieiga per internetą https://www.techtarget.com/searchdatamanagement/tip/Top-5-elements-needed-for-a-successful-data-warehouse
32. Kurunji, S., Ge, T., Fu, X., Liu, B. ir Chen, C.X. (2011) Optimizing Communication for MultiJoin Query Processing in Cloud Data Warehouses. Prieiga per internetą http://www.cs.uml.edu/~xinwenfu/paper/Journals/13_IJGHPC_5_4_OptimizingCommunicationforMulti-JoinQueryProcessinginCloudDataWarehouses.pdf
33. L'Heureux, A., Grolinger, K. ir Capretz, M.A.M. (2017) Machine Learning With Big Data: Challenges and Approaches. Prieiga per internetą https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7906512
34. Lebenthal, A. (2018) Data Architecture vs Information Architecture: What's The Difference? Prieiga per internetą https://www.bmc.com/blogs/data-architecture-vs-information-architecture/
35. Leščevskij, K. (2019). Viskas, ką turite žinoti apie įmonės veiklos rodiklius (KPI). https://www.cet.lt/viskas-ka-turite-zinoti-apie-imones-veiklos-rodiklius-kpi/
36. Malak, H.A. (2021) 11 Data Management Challenges and Solutions. Prieiga per internetą https://theecmconsultant.com/data-management-challenges/
37. Nargesian, F., Zhu, E., Miller, R.J ir Pu, K.Q. (2019) Data Lake Management: Challenges and Opportunities. Prieiga per internetą http://www.cs.toronto.edu/~fnargesian/Data_Lake_Management.pdf
38. Oketunji, T.A. ir Omodara, R.O. (2011) Design of Data Warehouse and Business Intelligence System. Prieiga per internetą https://www.diva-portal.org/smash/get/diva2:831050/FULLTEXT01.pdf
39. Oracle Database Online Documentation 12c Release. Introduction to Data Warehousing Concepts. Prieiga per internetąhttps://docs.oracle.com/database/121/DWHSG/concept.htm#DWHSG-GUID-452FBA23-6976-4590-AA41-1369647AD14D
40. Ordonez, C., Song, I.Y., Kotsis, G.A., Tjoa, A.M. ir Khalil, I. (2019) Big Data Analytics and Knowledge Discovery. Prieiga per internetą https://link.springer.com/content/pdf/10.1007/978-3-030-27520-4.pdf
41. Qualitative vs. Quantitatvie Research. Prieiga per internetą https://www.5staressays.com/blog/qualitative-vs-quantitative-research/qualitative-vs-quantitative-research.pdf

42. Riley, J. (2017). Understanding Metadata What is Metadata, and What is it for? Prieiga per internetą https://www.fidgeo.de/wp-content/uploads/2016/07/2017_01-NISO-understanding-metadata.pdf
43. Sample size calculator & complete guide (2022). Prieiga per internetą https://www.qualtrics.com/blog/calculating-sample-size/
44. Sarda, N.L. (2001) Structuring Business Metadata in Data Warehouse Systems for Effective Business Support. Prieiga per internetą https://arxiv.org/pdf/cs/0110020.pdf
45. Seymour, T.J. ir Berg, K. History of Databases. Prieiga per internetą https://www.researchgate.net/publication/298332910_History_Of_Databases
46. Singh, A. ir Ahmad, S. (2019) Architecture of Data Lake. Prieiga per internetą https://www.researchgate.net/profile/Ajit-Singh-46/publication/331890045_Architecture_of_Data_Lake/links/6061ef85458515e8347d6ecc/Architecture-of-Data-Lake.pdf
47. Spacey, J. (2018) 6 Examples of Analytical Data. Prieiga per internetą https://simplicable.com/new/analytical-data
48. Stedman, C. (2021) What is data architecture? A data management blueprint. Prieiga per internetą https://www.techtarget.com/searchdatamanagement/definition/What-is-data-architecture-A-data-management-blueprint
49. The Hadoop ecosystem. Prieiga per internetą https://www.ibm.com/analytics/hadoop
50. What is Lambda Architecture? Prieiga per internetą https://databricks.com/glossary/lambda-architecture

SANTRAUKA

VILNIAUS UNIVERSITETAS

EKONOMIKOS IR VERSLO ADMINISTRAVIMO FAKULTETAS

VERSLO INFORMACINIŲ SISTEMŲ PROGRAMA

MARIUS STULGINSKAS

DUOMENŲ TINKLO ARCHITEKTŪROS POREIKĮ LEMIANTYS VEIKSNIAI ORGANIZACIJOSE

Darbo vadovas: Dr. Asistentas Michail Kazimianec

Darbas parengtas: 2022m. Vilnius

Darbo apimtis: 60 puslapių

Lentelių skaičius darbe: 7 lentelės

Literatūros šaltinių skaičius: 50 vnt.

Darbo apibūdinimas: darbe buvo analizuojamos duomenų ir informacijos valdymo architektūros, duomenų valdymo iššūkiai ir skirtingos duomenų saugojimui ir analizei skirtos technologijų architektūros. Atliktu tyrimu buvo siekta išsiaiškinti, kokie yra veiksniai, kurie lemia duomenų tinklo architektūros (angl. Data Mesh) poreikį organizacijose. Taip pat tyrimu siekta išsiaiškinti ir kitus su duomenų valdymo iššūkiais susijusius veiksnius ir dažniausiai pasikartojančius jų derinius, atsižvelgiant į respondentų atsakymus. Tyrimo metu gauti rezultatai apdoroti naudojant IBM SPSS Modeler programinę įrangą.

Darbo tikslas: nustatyti esminius veiksnius, lemiančius duomenų valdymo problemas ir iššūkius

Darbo uždaviniai: Remiantis literatūros analize palyginti informacijos ir duomenų valdymo architektūras; Išnagrinėti klasikinę verslo įžvalgos ir duomenų valdymo architektūrą; Išnagrinėti kokią naudą organizacijai teikia efektyvus metaduomenų valdymas; Išnagrinėti su kokiomis pagrindinėmis problemomis susiduria organizacijos siekdamos efektyviai valdyti duomenis; Remiantis apklausos rezultatais, įvertinti įmonių pasirengimą ir poreikį pritaikyti duomenų tinklo architektūrą; Pateikti išvadas ir pasiūlymus/rekomendacijas.

SUMMARY

VILNIAUS UNIVERSITY

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

BUSINESS INFORMATION SYSTEMS PROGRAM

MARIUS STULGINSKAS

DRIVERS OF THE NEED FOR DATA MESH ARCHITECTURE IN ORGANIZATIONS

Supervisor: Dr. Assistant Michail Kazimianec

Bachelor thesis was written in: 2022m. Vilnius

Size: 60 pages

Number of tables: 7 tables

Number of referencees: 50

Description of the work: this work analysed the architectures of data and information management, data management challenges and different technology architectures, which are used for data storing and analysis. The study sought to find out what are the drivers of the need for Data Mesh architecture in organizations. It also sought to identify other factors related to data management challenges and the most recurrent combinatons of these factors, based on respondents answers. The results were processed using IBM SPSS Modeler software.

Aim of the work: identify main factors, which leads to data management problems and challenges

Work‘s objectives: Based on the literature analysis, compare information and data management architectures; Examine the classic business intelligence and data management architecture; Examine the benefits of effective metadata management for the organization; Examine the main challenges faced by organizations for effective data management; Based on the survey results, assess the readiness and need of companies to adopt a Data Mesh architecture; Provide conclusions and suggestions/recommendations.

PRIEDAI

1 priedas

Klausimynas/anketa

Duomenų valdymo iššūkiai organizacijoje

Gerb. Respondente,

Esu Vilniaus universiteto Ekonomikos ir verslo administravimo fakulteto IV kurso studentas Marius Stulginskas.

Kviečiu Jus sudalyvauti anketinėje apklausoje. Jūsų išsakyta nuomonė padės nustatyti, su kokiais pagrindiniais iššūkiais susiduria organizacijos ir darbuotojai tvarkant ir valdant duomenis.

Prašau atsakyti į visus anketos klausimus. Anketa yra anoniminė, o atsakymai bus nagrinėjami tik tyrimo analizei atlikti. Anketą užpildyti truksite 5-10min. Iš anksto dėkoju Jums.

Šios anketos rezultatai viešai nepublikuojami.

1. Kokią technologiją naudojate duomenų kaupimui, kuri įgalina analitiką Jūsų organizacijoje?

    a. Duomenų saugyklos technologija
    b. Duomenų ežero technologiją

2. Su kokiomis problemos susiduriate naudojant duomenų saugyklos (angl. Data Warehouse) technologiją? (1 - visiškai nesutinku, 5 - visiškai sutinku)

    a. Lėtas užklausų atsakas/vykdymas
    b. Duomenų modeliavimo trūkumas
    c. Automatizacijos trūkumas (rankiniai sistemos ir duomenų atnaujinimai)
    d. Trūksta aprašomųjų metaduomenų arba metaduomenų palaikymo mechanizmo, kuris užtikrintų kokybišką metaduomenų valdymą
    e. Prastas duomenų šaltinių atsekamumas
    f. Ribota prieiga prie duomenų
    g. Aktualių duomenų trūkumas reikiamu momentu
    h. Duomenys netikslūs arba jų trūksta
    i. Duomenų dubliavimasis
    j. Netikslūs duomenų laukų pavadinimai
    k. Sistemos ir procesų mokymų trūkumas naudotojams
    l. Duomenų darnos (integracijų) trūkumas
    m. Negarantuotas sistemos našumas
    n. Nėra vieningos duomenų saugyklos architektūros
    o. Nėra vieningo duomenų saugyklos valdymo sprendimo

3. Su kokiomis problemomis susiduriate naudojant duomenų ežero (angl. Data Lake) technologiją? (1 - visiškai nesutinku, 5 - visiškai sutinku)

    a. Sistema priima bet kokius duomenis be jų peržiūros ar valdymo
    b. Trūksta aprašomųjų metaduomenų arba metaduomenų palaikymo mechanizmo, kuris užtikrintų kokybišką metaduomenų valdymą
    c. Duomenis reikia nuolat analizuoti iš naujo
    d. Negarantuotas sistemos našumas
    e. Skirtingi duomenų formatai
    f. Duomenų priėmimo realiu laiku trūkumas
    g. Spartesnio duomenų išgavimo trūkumas
    h. Nėra vieningos duomenų ežero architektūros
    i. Nėra vieningo duomenų ežero valdymo sprendimo

1. Su kuriomis žemiau pateiktomis problemomis susiduriate valdant ir tvarkant duomenis ir su kuriais teiginiais sutinkate? (1 - visiškai nesutinku, 5 - visiškai sutinku)

    a. Sunkus duomenų savininkų atsekamumas
    b. Sunkus duomenų šaltinių atsekamumas
    c. Sudėtinga apibrėžti savo naudotojų ratą
    d. Sudėtinga sudaryti tikslius įmonės pajėgumo planus, siekiant konkrečių tikslų
    e. Duomenų architektai/informacijos modeliuotojai/duomenų analitikai susiduria su duomenų siloso problema
    f. Egzistuoja atskiri duomenų domenai
    g. Duomenys organizacijoje traktuojami kaip turtas (angl. data as an asset)
    h. Duomenys organizacijoje traktuojami kaip produktas (angl. data as a product)
    i. Aprašomi duomenų loginiai modeliai
    j. Prieigos kontrolės spraga (duomenys gali būti pakeisti be turinio peržiūros arba be duomenų savininko žinios)
    k. Didelė dalis verslui svarbios informacijos pateikiama per vėlai
    l. Nepakankami sistemos ištekliai, didėjančiam duomenų kiekiui apdoroti
    m. Pilnai neišnaudojamas duomenų potencialas, maksimaliai naudai pasiekti
    n. Duomenų kiekis auga greičiau, nei tobulėja technologinė infrastruktūra jiems apdoroti
    o. Sudėtinga atrinkti reikiamus metaduomenis
    p. Veiklos procesų automatizavimo trūkumas
    q. Prieiga prie jautrių asmens duomenų yra apribota ir reguliariai stebima
    r. Naudojami duomenų analitikos įrankiai
    s. Organizacijoje atskiri padaliniai atsakingi už savo valdomus duomenis ir už su jais susijusius sprendimus (angl. Ownership and accountability)
    t. Nustatyti duomenų kokybės standartai
    u. Aiškiai išskirtos rolės ir atsakomybės, skirtingiems duomenų valdymo komponentams (duomenų planavimas, duomenų procesai, duomenų sauga ir t.t.)

v. Naudojamos(-a) duomenų kokybės valdymo ir priežiūros sistemos(-a) (angl. Data governance)

2. Jūsų amžius:

    a. 18-25 metai
    b. 26-35 metai
    c. 36-45 metai
    d. 46-55 metai
    e. 56 ir daugiau

3. Koks Jūsų išsilavinimas? (pasirinkite tik vieną atsakymą)

    a. Pradinis
    b. Pagrindinis (nebaigtas vidurinis)
    c. Vidurinis
    d. Profesinis (profesinė mokykla, vidurinis su profesine kvalifikacija)
    e. Aukštesnysis (technikumas, aukštesniosios mokyklos)
    f. Neuniversitetinis aukštasis (kolegija)
    g. Universitetinis aukštasis (bakalauro laipsnis)
    h. Universitetinis aukštasis (magistro laipsnis, ar jam prilygstanti profesinė kvalifikacija/mokslų daktaras)

4. Kiek laiko dirbate šioje organizacijoje?

    a. Taip
    b. Ne

5. Kiek darbuotojų dirba Jūsų organizacijoje?

    a. Mažiau už 250
    b. Daugiau arba lygu 250

6. Jūsų organizacijos veiklos sritis?

    a. Žemės ūkis, žuvininkystė
    b. Energetika
    c. Statyba
    d. Gamyba

e. Didmeninė ir mažmeninė prekyba
f. Transportas
g. Apgyvendinimo ir maitinimo paslaugos
h. Informacija ir ryšiai, IT
i. Finansinė ir draudimo veikla
j. Paslaugų verslas
k. Švietimas, mokslas
l. Sveikatos priežiūra
m. Kultūra
n. Kita