Duomenų tyrybos konspektas

2. Apie duomenis išsamiau

2.1. Didieji duomenys (Big Data)

Svarbiausios didžiųjų duomenų charakteristikos

Šaltinis	Svarbiausios charakteristikos
Wang, J., Xu, C., Zhang, J., & Zhong, R. (2022). Big data analytics for intelligent manufacturing systems: A review. *Journal of Manufacturing Systems*, *62*, 738-752.	Sparčiai vystantis šiuolaikinėms technologijoms daugiausia dėmesio skiriama milžiniškų duomenų kiekių, vadinamų didžiaisiais duomenimis, rinkimui. Tačiau susiduriama su dideliais iššūkiais apdorojant didžiuosius duomenis ir išgaunant iš jų informaciją. Didieji duomenys apibūdinami "3V" charakteristikomis – apimties (Volume), įvairovės (Variety) ir greičio (Velocity).
Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2019). Big data adoption: State of the art and research challenges. *Information Processing & Management*, *56*(6), 102095.	Atsiradus kompiuteriams, internetui ir mobiliosioms technologijoms duomenų kiekiai sparčiai auga. Duomenys generuojami per socialinę žiniasklaidą, apsipirkimą internetu, sandorius, tinklo įrenginius ir švietimo įrašus. Anksčiau duomenys buvo saugomi duomenų bazėse ir skaičiuoklėse, o informacija, kurią sunku konvertuoti į eilutes ir stulpelius, buvo ignoruojama. Dideli duomenys yra įvairių tipų: struktūrizuoti (duomenų bazės, sukurtos naudojant SQL serverį ir Oracle ir kt.), nestruktūrizuoti (vaizdo įrašai, garso įrašai, įvairūs dokumentai, vaizdai, komentarai, stebėtojai, patinkantys, žymos, tviteriai, paspaudimai ir pokalbiai ir kt.) ir pusiau struktūrizuoti (trečiųjų šalių duomenys, valiutos konvertavimas, orai, XML, grafiko ar teksto duomenys, elektroninė prekyba ir kt.). Didelių duomenų paslaugos suteikia galimybes apdoroti, analizuoti įvairių tipų duomenis ir pateikti prognozes per trumpą laiką.
Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, *30*(4), 431-448.	Skirtingai nuo tradicinių duomenų, terminas „didieji duomenys“ apibūdina didelius heterogeninius (nevienalyčius) duomenų rinkinius (nuolat didėjančius), apimančius struktūrizuotus, nestruktūrizuotus ir pusiau struktūrizuotus duomenis. Didieji duomenys yra sudėtingi, todėl jiems apdoroti reikalingos galingos technologijos ir pažangūs algoritmai. Dauguma duomenų mokslininkų ir ekspertų didžiuosius duomenis apibrėžia trimis pagrindinėmis charakteristikomis (vadinamos 3V): Apimtis (Volume): nuolat generuojami dideli skaitmeninių duomenų kiekiai iš milijonų įrenginių ir programų (IRT, išmaniųjų telefonų, produktų kodai, socialiniai tinklai, jutikliai, žurnalai ir kt.). Greitis (Velocity): duomenys generuojami greitai ir turi būti apdorojami greitai norint gauti naudingos ir svarbios informacijos. Įvairovė (Variety): dideli duomenys generuojami iš paskirstytų įvairių šaltinių ir įvairiais formatais (pvz., vaizdo įrašais, dokumentais, komentarais, jutiklių duomenys). Dideli duomenų rinkiniai gali būti vieši ar privatūs, vietiniai ar nuotoliniai, bendrinami arbakonfidencialūs, išsamūs ar neišsamūs ir pan.

Emani, C. K., Cullot, N., & Nicolle, C. (2015). Understandable big data: a survey. *Computer science review*, *17*, 70-81.	Didieji duomenys – tai duomenų rinkiniai, kurių dydis viršija tipinių duomenų bazių programinės įrangos įrankių galimybes duomenis fiksuoti, saugoti, tvarkyti ir analizuoti. Taip pat akcentuojamas duomenų daugiamatiškumas (multi-dimensionality), greitas judėjimas (moves to fast). Didžiųjų duomenų charakteristikos papildytos dar 3V: Vertė (Value) - duomenų architektūros yra sukurtos ekonomiškai išgauti vertę iš labai didelių įvairių duomenų kiekių, įgalinant didelio greičio fiksavimą, atradimą ir (arba) analizę“. Vertė skirstoma į dvi kategorijas: analitiniam naudojimui (pakeitimui / palaikymui žmogaus sprendimas, poreikių atradimas, populiacijų segmentavimas pritaikyti veiksmus) ir naujų verslo modelių, produktų ir paslaugų įgalinimui. Duomenų teisingumas, tikslumas, tikrumas (Veracity) – Netikslumus gali sukelti nepilni duomenys, modelio tikslumas, neaiškumai duomenyse, apgaulė, sukčiavimas, dubliavimas, neužbaigtumas, šlamštas ir vėlavimas. Dėl teisingumo rezultatai gauti iš didelių duomenų, negali būti įrodyti. Jie yra tikimybiniai (pateikiami su tam tikra tikimybe). Vizualizacija (Visualization) – reikalinga analizės rezultatų pateikimui. Apibendrinant galima teigti, kad norint veiksmingai dirbti su didžiaisiais duomenimis reikia kurti vertę (Value), atsižvelgiant į duomenų apimtį (Volume), įvairovę (Variety) ir teisingumą (Veracity), kol jie vis dar juda (Velocity), o ne tik tada, kai jie jau yra (at the rest).
Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*.	Nepaisant to, kad egzistuoja įvairūs didžiųjų duomenų apibrėžimai, visuose galima išskirti šiuos pagrindinius veiksnius: Dydis: duomenų rinkinių apimtis. Sudėtingumas: duomenų struktūra, elgsena ir kombinacijos. Technologijos: priemonės ir metodai, kurie naudojami apdorojant didelius ar sudėtingus duomenų rinkinius. Suformuluotas apibrėžimas: didieji duomenys apima didelės apimties sudėtingų duomenų rinkinių saugojimą ir analizę panaudojant įvairias technologijas.
Yu, S., Liu, M., Dou, W., Liu, X., & Zhou, S. (2016). Networking for big data: A survey. *IEEE Communications Surveys & Tutorials*, *19*(1), 531-549.	Tai duomenų rinkiniai, kurių dydžiai viršija įprastai naudojamų programinės įrangos priemonių galimybes, tvarkyti ir apdoroti duomenis per „leistiną laiko tarpą“.
Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.	Išskiriamos svarbiausios didžiųjų duomenų savybės: heterogeniškumas (nevienalytiškumas), triukšmų kaupimo efektas, klaidinga koreliacija, atsitiktinis endogeniškumas (kintamojo priklausomybė nuo kitų sistemos kintamųjų, https://lt1.uppercreditfieldnaturalists.org/what-is-endogeneity-what-is-an-exogenous-variable-2a66c4b). Šios savybės neleidžia panaudoti tradicinių statistinių duomenų analizės metodų.

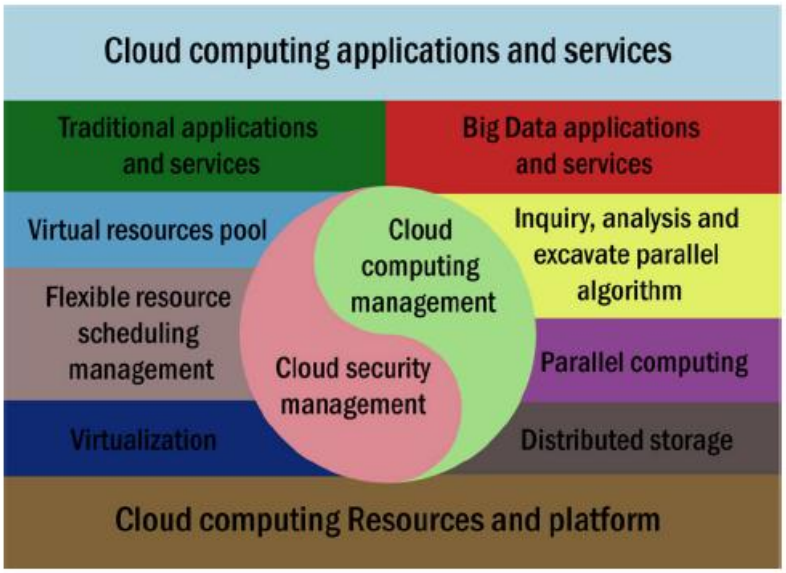
Abdalla, H. B. (2022). A brief survey on big data: technologies, terminologies and data-intensive applications. *Journal of Big Data*, *9*(1), 1-36.	Didieji duomenys išgaunami iš skirtingos kilmės ir įvairių konfigūracijų šaltinių (pvz., įrašai, ataskaitos, pastabos ir žurnalai). Didžiųjų duomenų rinkinius sudaro duomenys, kurie yra organizuoti ir nesutvarkyti, privatūs ir vieši, lokalūs ir nutolę, paskirstytieji ir slapti, pilni ir nepilni. Dauguma tyrėjų didžiuosius duomenis apibūdina penkiais atributais, vadinamais 5V: apimtis (Volume), greitis (Velocity), įvairovė (Variety), teisingumas (Veracity) ir vertė (Value).
Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*, 171-209.	Didieji duomenys apibūdina duomenų rinkinius, kurių negalima suvokti, įgyti, valdyti ir apdoroti naudojant tradicinius IT ir programinės/aparatinės įrangos įrankius per toleruotiną laiką. Atsižvelgiant į skirtingus aspektus, mokslo ir technologijų įmonės, mokslininkai, duomenų analitikai, technikos specialistai skirtingai apibrėžia didžiuosius duomenis. Išskiriamos pagrindinės didžiųjų duomenų savybės: 1) didelė apimtis ir įvairovė, sudėtinga struktūra; 2) duomenų išteklių našumas, kuris pasiekiamas pertvarkant ir integruojant iš skirtingų išteklių išgautus duomenis, kurie gali sukurti didesnę vertę; 3) didieji duomenys skatina kryžminį mokslo susiliejimą (cross fusion of science) apimantį ne tik debesų kompiuteriją, daiktų internetą, duomenų centrus ir mobiliuosius tinklus, bet ir duomenų išgavimo, saugumo ir analizės naujų technologijų ir metodų kūrimą; 4) vizualizacija, skirta efektyviam rezultatų panaudojimui; 5) orientuota į duomenis (data-oriented) programos architektūra pakeičia į algoritmus orientuotą architektūrą; 6) didžiųjų duomenų analizės rezultatai padės priimti tinkamesnius sprendimus.
https://www.europarl.europa.eu/news/lt/headlines/society/20210211STO97614/didieji-duomenys-apibrezimas-nauda-issukiai-infografikas#:~:text=Didieji%20duomenys%20%28angl.%20big%20data%29%20%E2%80%93%20tai%20%C5%BEmoni%C5%B3,sistem%C5%B3%2C%20pvz.%2C%20GPS%2C%20signalai%2C%20jutikli%C5%B3%20registruojama%20klimato%20informacija.	Didieji duomenys (angl. *big data*) – tai žmonių ar mašinų sukuriami dideli duomenų kiekiai, kaip antai pirkimo sandorių duomenys, naudojimosi socialiniais tinklais įpročiai, interesai ir pomėgiai, padėties nustatymo sistemų, pvz., GPS, signalai, jutiklių registruojama klimato informacija. Apdoroti ir išanalizuoti jie paverčiami įžvalgomis, padedančiomis priimti sprendimus rinkodaros, reklamos, prekybos, miestų planavimo, sveikatos priežiūros bei transporto srityse. Pasinaudodamos šia informacija įmonės gali siūlyti prekes tik tiems, kam jų reikia, ir tik tada, kai jų reikia. Ji taip pat gali patarti, kokį maršrutą pasirinkti, kokią draudimo paslaugą pasiūlyti klientui, ar suteikti paskolą ir netgi gali padėti išsirinkti tinkamą darbuotoją.

Didžiųjų duomenų sąsajos su šiuolaikinėmis technologijomis

Šaltinis: Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*, 171-209.

Didieji duomenys yra glaudžiai susiję su **debesų kompiuterija** (Cloud Computing), nes jų saugojimui reikia didelių saugyklų, o apdorojimui ir analizei – daug skaičiavimų. Iš kitos pusės,

didieji duomenys spartina debesų kompiuterijos vystymąsi, kuriant paskirstytąsias saugyklas ir lygiagrečiuosius skaičiavimus, skirtus didiesiems duomenims.

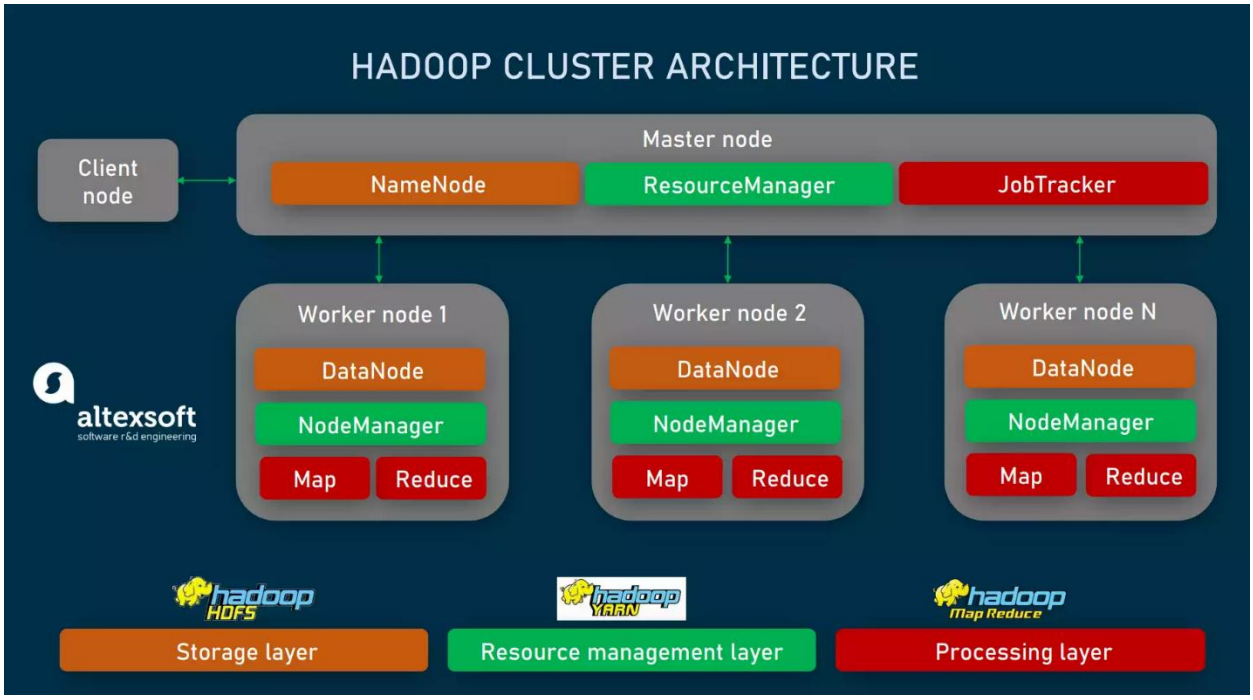


Daiktų internetas (Internet of Things) pasižymi tuo, kad jutikliai yra įmontuoti į įvairius įrenginius ir mašinas realiame pasaulyje. Tokie jutikliai naudojami įvairiose srityse gali rinkti įvairių rūšių duomenis: aplinkos, geografinius, astronominius, logistikos ir pan. Didieji daiktų interneto generuojami duomenys palyginus su bendrais didžiaisiais duomenimis, skiriasi renkamų duomenų tipais, yra heterogeniniai, nestruktūrizuoti, triukšmingi, pasikartojantys (besidubliuojantys). Daiktų interneto diegimas skatina spartų duomenų augimą pagal apimtį ir įvairovę. Didžiųjų duomenų technologijų perkėlimas į daiktų internetą spartina daiktų interneto pažangą ir verslo modelių kūrimą.

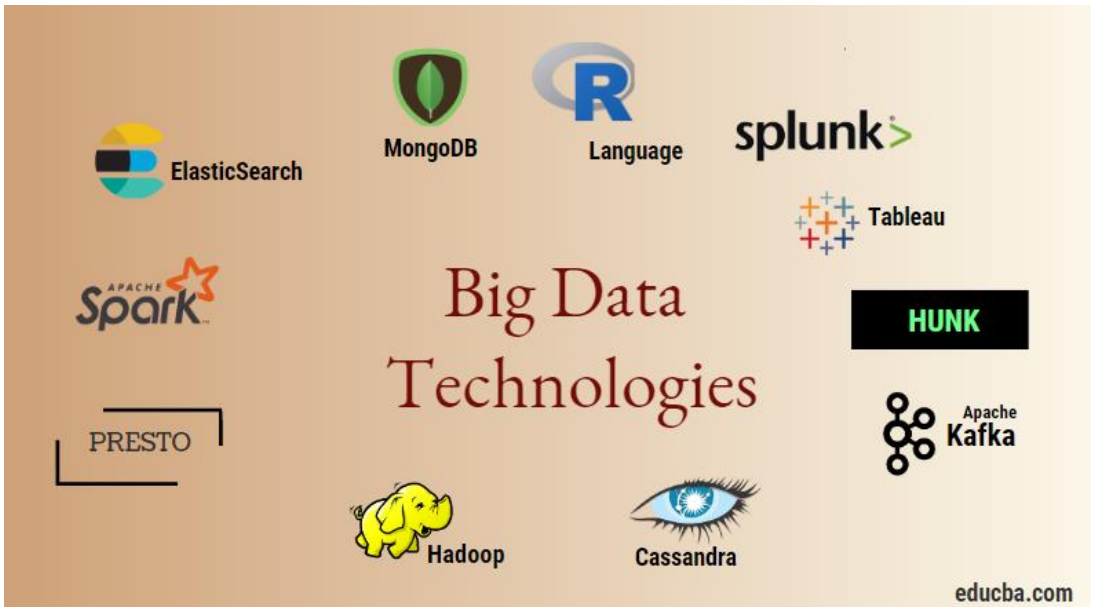
Didžiųjų duomenų paradigmoje **duomenų centras** (Data Center) yra ne tik platforma koncentruotam duomenų saugojimui, bet ir atlieka duomenų gavimo, tvarkymo, panaudojimo funkcijas. Dideliems duomenims reikia, kad duomenų centras užtikrintų saugojimo talpos, apdorojimo, tinklo perdavimo pajėgumus. Duomenų centras apima infrastruktūrą su daugybe mazgų, didelės spartos vidiniais tinklais, efektyviai išsklaido šilumą, daromos atsarginės duomenų kopijos. Didžiųjų duomenų taikymų plėtra spartina naujovių diegimą duomenų centruose. Toliau augant duomenų apimtims, struktūrizuoti ir nestruktūrizuoti duomenys bei jų įvairovė, analitinių duomenų šaltiniai, duomenų apdorojimas ir skaičiavimai tiesiogiai susiję su duomenų centro pajėgumų didinimu. Duomenų centrai apima ne tik techninę, bet ir programinę įrangą, skirtą didžiųjų duomenų organizavimui, analizei ir taikymui.

Šiuo metu **„Hadoop“** [1] plačiai naudojama didelių duomenų taikymuose: pvz., šiukšlių filtravimas, paieška tinkle, paspaudimų srauto analizė ir socialinės rekomendacijos. „Hadoop“ leidžia lengviau išnaudoti visą saugyklos ir apdorojimo talpą klasterių serveriuose ir vykdyti paskirstytus procesus, apimant didelius duomenų kiekius. „Hadoop“ ekosistema bėgant metams labai išaugo dėl savo išplėtimo. Šiandien Hadoop ekosistemoje yra daug įrankių ir programų, padedančių rinkti, saugoti, apdoroti, analizuoti ir valdyti didelius duomenis.

[1] „Hadoop“ yra atvirojo kodo paskirstytojo apdorojimo karkasas (framework), valdantis didžiųjų duomenų apdorojimą ir saugojimą keičiamo dydžio kompiuterių serverių grupėse. Tai didžiųjų duomenų technologijų, kurios pirmiausia naudojamos duomenų mokslo ir pažangios analizės iniciatyvoms, įskaitant prognostinę analizę, duomenų gavybą, mašininį ir gilųjį mokymąsi, ekosistemos centras (https://www.techtarget.com/searchdatamanagement/definition/Hadoop).



Šaltinis: https://www.altexsoft.com/blog/hadoop-vs-spark/

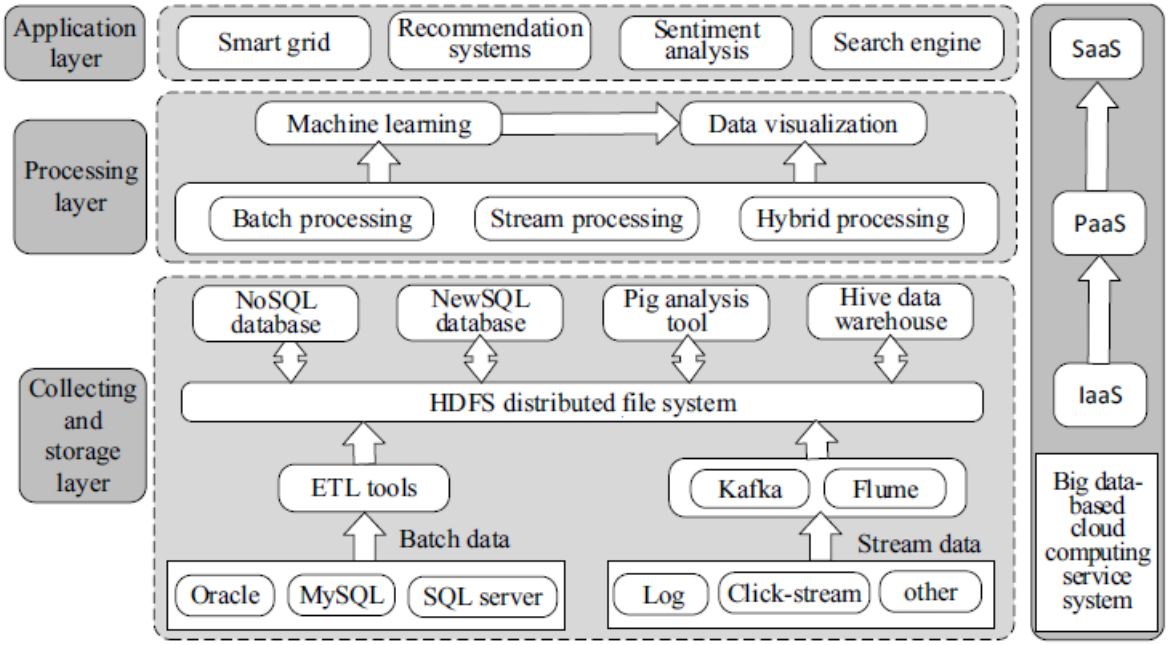


Šaltinis: https://www.educba.com/big-data-technologies/

Apie didžiųjų duomenų technologijas išsamiai rasite: Abdalla, H. B. (2022). A brief survey on big data: technologies, terminologies and data-intensive applications. *Journal of Big Data*, *9*(1), 1-36.

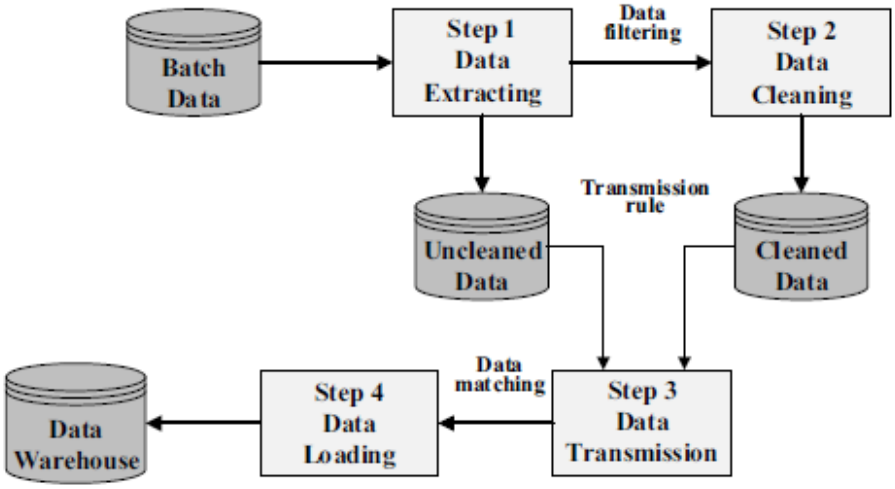
Didžiųjų duomenų paslaugų architektūra

Šaltinis: Wang, J., Yang, Y., Wang, T., Sherratt, R. S., & Zhang, J. (2020). Big data service architecture: a survey. *Journal of Internet Technology*, *21*(2), 393-405.



Architektūrą sudaro 3 sluoksniai. **Duomenų surinkimo ir saugojimo sluoksnyje** (Collecting and storage layer) duomenys surenkami panaudojant atitinkamą įrangą ir iš anksto apdoroti duomenys (pre-processed data) saugomi paskirstytoje failų sistemoje (HDFS (Hadoop distributed file system) distributed file system) arba duomenų bazėje.
Duomenų apdorojimo sluoksnyje (Processing layer) įvairūs duomenų apdorojimo karkasai pritaikomi skirtingiems duomenų tipams. Išsami duomenų analizė pagrįsta didelės apimties mašininio mokymosi technologijomis, kurios padidina duomenų vertę. Vizualizacijos įrankiai naudojami rezultatams pristatyti.
Taikymo sluoksnyje (Application layer) yra didžiųjų duomenų taikymai įvairiose srityse. Taip pat panaudojama debesų kompiuterija, kuri apima skaičiavimams skirtą programinę įrangą ir paslaugas (SaaS (Software as a Service) - programinės įrangos pristatymo ir licencijavimo būdas, kai programinė įranga pasiekiama internetu per prenumeratą, o ne perkama ir įdiegiama atskiruose kompiuteriuose, PaaS (Platform as a service) – yra debesų kompiuterijos modelis, suteikiantis klientams visą debesų platformą – aparatinę įrangą, programinę įrangą ir infrastruktūrą, skirtą programoms kurti, paleisti ir valdyti be sąnaudų, sudėtingumo ir nelankstumo, kurie dažnai atsiranda kuriant ir palaikyti tą platformą, IaaS (infrastructure as a Service) – tai yra debesų kompiuterijos forma, teikianti pagrindinius skaičiavimo, tinklo ir saugojimo išteklius vartotojams pagal poreikį, internetu ir už mokestį „you-go“ pagrindu. IaaS suteikia galimybę galutiniams vartotojams pagal poreikį padidinti ir sumažinti išteklius, todėl sumažėja didelių išankstinių kapitalo išlaidų arba nereikalingos „priklausomos“ infrastruktūros poreikis, ypač esant „smagiam“ darbo krūviui. Priešingai nei PaaS ir SaaS, IaaS įgalina žemiausio lygio išteklių valdymą debesyje.

Išsamiau aptarsime ETL (Extract-Transform-Load) – duomenų išgavimo, transformavimo ir įkėlimo procesą.

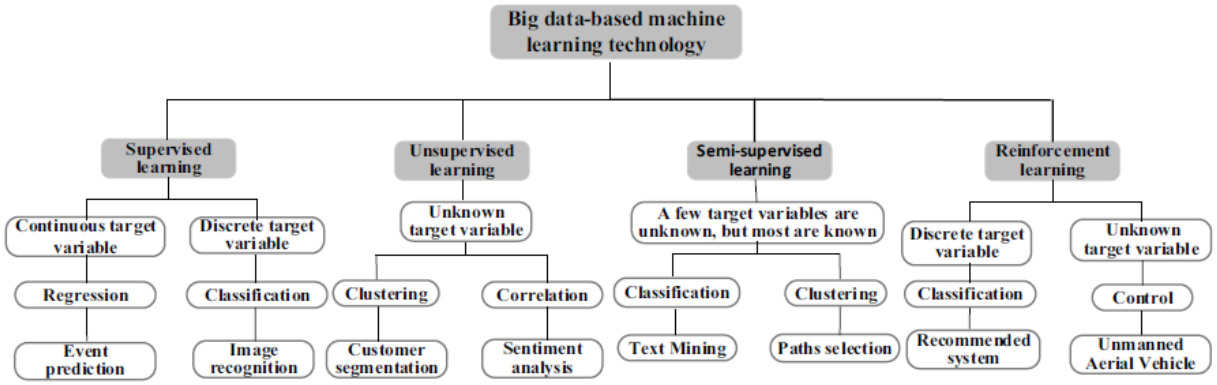


Didžiųjų duomenų formos apima **statinius paketinius duomenis** (Batch data, tai duomenų rinkinys arba grupė, kurią galima apdoroti arba analizuoti kaip vienetą) ir **dinaminius srauto duomenis** (Stream data, tai nuolatinis, besikeičiantis duomenų srautas). Dėl srauto duomenų perdavimo nestabilumo, jų rinkimas skiriasi nuo paketinių duomenų rinkimo. Paketinių duomenų surinkimui ir perdavimui (Transmission) iš skirtingų šaltinių naudojami ETL įrankiai.

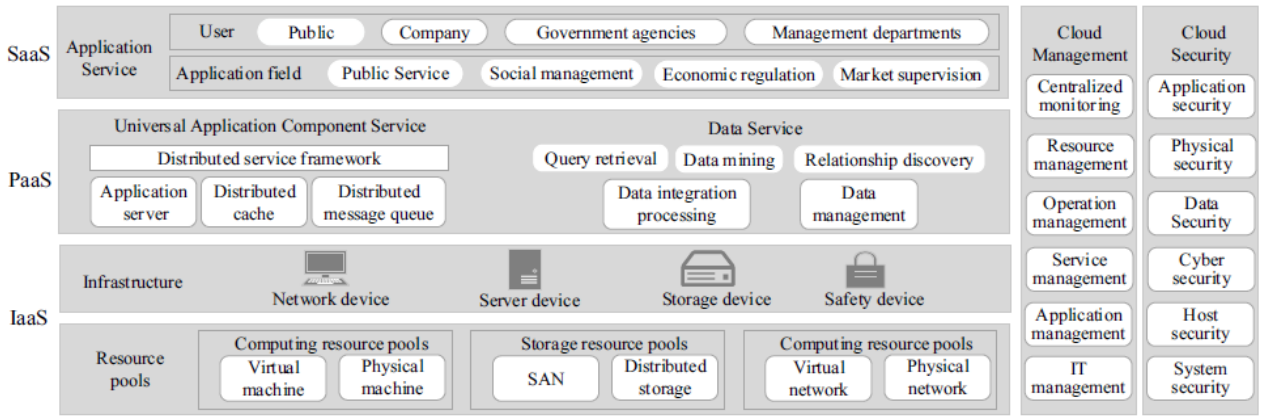
ETL pašalina sugadintus ar triukšmingus duomenis atlikdami duomenų apdorojimo operacijas, pvz., prijungimą, transformavimą ir valymą. Plačiai naudojami ETL įrankiai yra Kettle, Datastage, Informatica ir kt.

Srauto duomenims, kurie renkami realiame laike, reikalingi įrankiai, kurie garantuoja momentiškumą, atsparumą gedimams, stabilumą ir patikimumą. Flume yra patikima ir atspari gedimams paskirstyta srauto apdorojimo sistema, kuri renka, kaupia ir perduoda daug žurnalo duomenų iš skirtingų šaltinių į centralizuotą saugyklą. Kafka yra universali atviro kodo pranešimų sistema, kuri daugiausia naudojama kuriant realiu laiku duomenų srautus ir srautinio perdavimo taikymus. Siekiant dar labiau optimizuoti valdymą ir duomenų srauto apdorojimo greitį, Kafka apdorojimui naudoja eiles, kad būtų išvengta apdorojimo asinchroniškumo tarp duomenų generavimo ir apdorojimo greičio.

Apdorojant didžiuosius duomenis naudojamos **mašininio mokymosi [2]technologijos**.

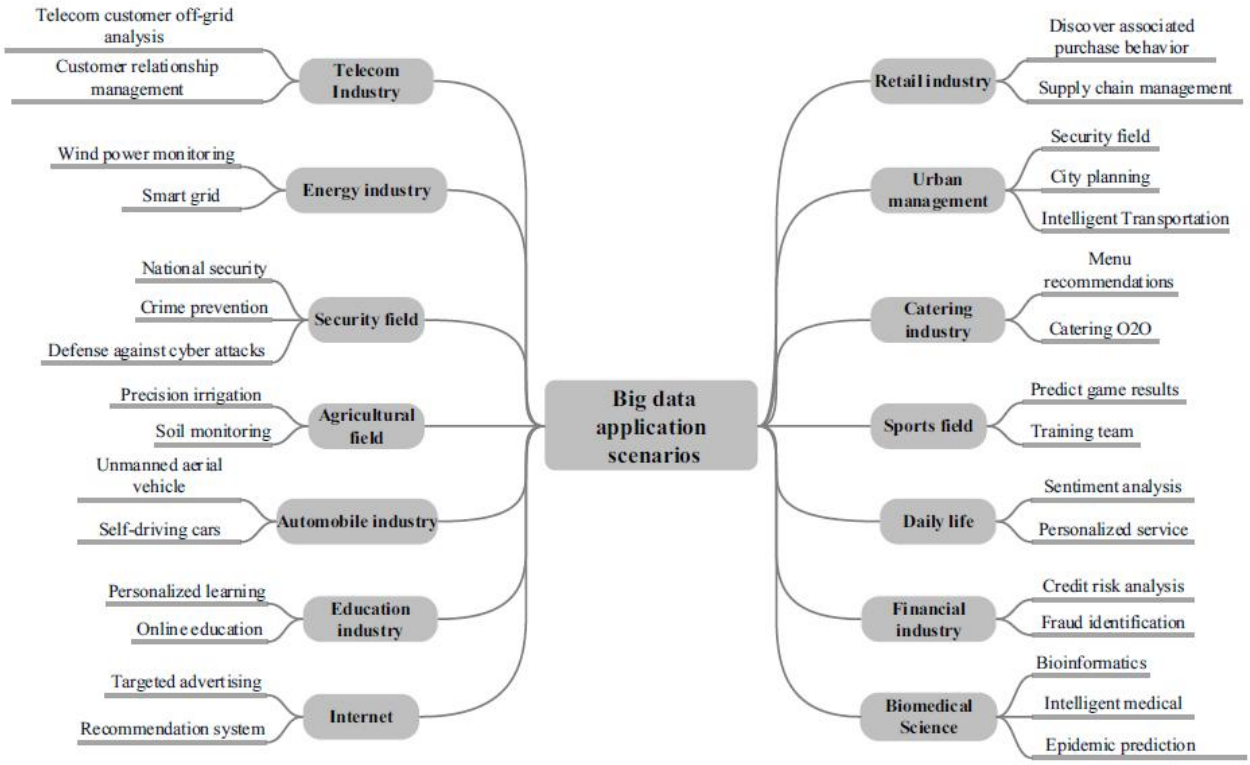


Didžiaisiais duomenimis pagrįstos **debesų kompiuterijos paslaugų sistemos** pasižymi tuo, kad turi galingus paskirstytus apdorojimo variklius, paskirstytas duomenų bazes, debesų saugyklą ir palaiko virtualizacijos technologijas.



[2] Mašininis mokymasis - dirbtinio intelekto algoritmų klasė, kuriai būdingas ne tiesioginis problemos sprendimas, o mokymasis, kaip pritaikyti daugelio panašių problemų sprendimus. Apima metodų, mokinančių kompiuterius „mąstyti“, kūrimą. Tai yra programų kūrimo būdas, kai sukurta sistema prisitaiko prie duomenų („apsimoko“). Šie algoritmai sugeba ilgainiui pasiekti geresnių rezultatų, patys kaupdami patyrimą.
Šaltiniai: https://lt.wikipedia.org/wiki/Ma%C5%A1ininis_mokymasis, A. Paulauskaitė-Tarasevičienė, K. Šutienė (2022). Intelektikos pagrindai, KTU leidykla „Technologija“

Didžiųjų duomenų taikymo scenarijai



Didžiųjų duomenų keliami iššūkiai

Šaltinis: Abdalla, H. B. (2022). A brief survey on big data: technologies, terminologies and data-intensive applications. *Journal of Big Data*, *9*(1), 1-36.

Iššūkiai, susiję su didžiųjų duomenų taikymais

- Dauguma didelių duomenų analizės sistemų apsiriboja „Apache Hadoop“ rinkiniu. Vadinasi, duomenų saugojimo ir valdymo sluoksnių išskyrimas yra iššūkis. Sluoksnių išskyrimas padidina didžiųjų duomenų funkcionalumą taikymuose.
- Pagrindinė problema, su kuria susiduriama, yra paskirstytų ir centralizuotų funkcijų derinimas didžiųjų duomenų taikymuose, pagrįstuose **giliuoju mokymusi** [3](deep learning).

Iššūkiai, susiję su didžiųjų duomenų analitika

- Duomenų apsauga, saugumas ir pasitikėjimas. Labai aiškiai apibrėžti duomenų naudojimo tikslai, asmens duomenų apsauga.
- Dalijimasis informacija ir jos valdymas. Ypatingai daug dėmesio skiriama atviriesiems duomenims, jų naudojimui laikantis įstatymų.
- Inovacijos ir analitika: saugyklos, duomenų išgavimas iš saugyklų; duomenų kiekio didėjimas; struktūrizuoti ir nestruktūrizuoti duomenys; duomenų nuosavybės ir iškraipymo problemos; duomenų įvairovė; riboti resursai didžiųjų duomenų apdorojimui; ribinių duomenų (edge data) apdorojimas; energijos suvartojimas.

Šaltinis: Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.

[3] Gilusis mokymasis yra mašininio mokymosi atšaka, kur kiekvienas sluoksnis išmoksta vis abstraktesnių duomenų bruožų. Tai neuroninis tinklas, turintis daugybę sluoksnių. Šaltinis: A. Paulauskaitė-Tarasevičienė, K. Šutienė (2022). Intelektikos pagrindai, KTU leidykla „Technologija“

- Sudėtingi duomenys, išgaunami iš įvairių šaltinių. Sunku numatyti priklausomybes tarp šių duomenų.
- Triukšmingi duomenys: matavimo paklaidos, nuokrypiai, trūkstamos reikšmės.
- Priklausomi duomenys: dalis duomenų yra susiję su santykinai silpnais signalais.

Šaltinis: Rakšnys, A. V., Gudelis, D., & Guogis, A. (2021). Didžiųjų duomenų ir dirbtinio intelekto technologijų pritaikymo galimybių viešojo valdymo srityje ir socialinėje politikoje analizė.

Dėl savo turimų įsipareigojimų viešojo sektoriaus institucijos, pradedant sveikatos apsauga ir socialinėmis paslaugomis, surenka daugybę jautrios informacijos apie savo piliečius. Šiuos duomenis naudojant kartu, būtų galima susidaryti beveik pilną individo gyvenimo vaizdą, tokiu būdu kėsinantis į asmeninio gyvenimo privatumą. Organizacijos gali rinkti duomenis tik tada, kai to reikia įgyvendinti jų misiją, tačiau toks „informacinis delegavimas“ dažnai yra neaiškus, nestabilus ir gali būti įvairiai interpretuojamas.

Vertinant iš etinės perspektyvos, dėl vartotojų duomenų panaudojimo galėtų būti labai sudėtinga arba net neįmanoma institucijoms gauti paslaugų vartotojų sutikimą. Dar vienas klausimas kyla dėl to, kad net jeigu ir pavyktų sutikimą gauti, tai ar ir kiek vartotojai galėtų būti informuojami apie jų duomenų panaudojimą (pvz., suprantant, kam ir kaip jie bus naudojami), nes labai dažnai didieji duomenys gali būti renkami turint vieną tikslą, o panaudojami visai kitam tikslui (Gillingham and Graham, 2016). Yra grėsmė, kad, turint „didžiuosius duomenis“, gali būti peržengta „raudonoji linija“ vartotojų atžvilgiu, pavyzdžiui, pasitaikius klaidų, kai į rinką rinkodaros būdu būtų prastumiami kenksmingi jiems produktai (Coulton et. al., 2015). Didieji duomenys yra susiję ir su pačių piliečių socialiniu aktyvumu. E. Hargittai (2015) atkreipia dėmesį į situacijas, kai pasikliovimas tik didžiaisiais duomenimis gali sukurti papildomų sunkumų, kai tam tikri svarbūs reiškiniai gali likti nepastebėti, nes jie susiję su tuo gyventojų segmentu, kuris nepalieka reikšmingo skaitmeninio pėdsako. Vyresni žmonės ir individai, gaunantys mažesnes pajamas, taip pat – individai, stokojantys išsilavinimo, nėra tinkamai reprezentuojami socialinės medijos duomenyse, o jaunimas ir aukštesniam socialiniam sluoksniui priklausantys individai yra per daug atspindimi duomenimis apie viešųjų paslaugų vartojimą (Coulton et. al., 2015). Veiksmingai naudoti didžiuosius duomenis būtinos ir tokios specifinės valstybės tarnautojų kompetencijos: programavimo, modeliavimo, statistikos, duomenų valdymo, analitikos įgūdžiai, mašininio mokymosi procesų supratimas, taip pat technologinė infrastruktūra (Maciejewski, 2016).

Analizuojant esminius trūkumus, svarbu paminėti ir pačių duomenų skaitmenizavimo procesų sudėtingumą, skirtingą duomenų pobūdį ir su jais susijusį netikrumą. Institucijos gali susidurti su iššūkiais numatydamos duomenų saugojimo vietų kūrimą, duomenų sistemų, kurios galėtų duomenis sujungti iš skirtingų šaltinių, pvz., kitų viešojo sektoriaus institucijų ar net nevyriausybinių bei verslo struktūrų, surinkimo kūrimą ir įdiegimą. Paminėtinas ir veiksmingas duomenų analizės metodų ir technikų poreikis, siekiant užtikrinti kompleksinį duomenų valdymą (Rogge et al., 2017). Dirbant su didžiaisiais duomenimis, tradicinių duomenų valdymo metodų nepakanka (Coulton et. al., 2015).

Skyrelio 2.1. medžiagai įtvirtinti siūlomas testas su pasirenkamaisiais atsakymais (anglų k.)

Šaltinis: https://study.com/academy/exam/topic/big-data-fundamentals.html

Alternatyvos: minčių žemėlapio sudarymas, papildomi didžiųjų duomenų atributai remiantis straipsniu: Khan, N., Naim, A., Hussain, M. R., Naveed, Q. N., Ahmad, N., & Qamar, S. (2019, May). The 51 v's of big data: survey, technologies, characteristics, opportunities, issues and challenges. In *Proceedings of the international conference on omni-layer intelligent systems* (pp. 19-24).