Tehnologija velikih količina podataka i digitalni arhivi

Šušnjara, Borna

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2021./2022.

Borna Šušnjara

Tehnologija velikih količina podataka i digitalni arhivi

Završni rad

Mentor: dr. sc. Arian Rajh, nasl. izv. prof.

Zagreb, lipanj 2022.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

Sadržaj

1. Uvod

„Big data“ je pojam koji postoji već dugo vremena, no posljednjih godina te kroz cijelo 21.st. postaje fenomen koji se proširio na razna područja te obuhvaća sve više poslovnih i znanstvenih polja u svom utjecaju. „Big data“ ili velike količine podataka nusprodukt su tehnološke i informacijske revolucije današnjeg doba. Sva područja poslovanja i svakodnevnog života doprinose rastućoj hrpi velikih podataka: maloprodaja, nekretnine, turizam, financije, društveni mediji te tehnologije, svaki aspekt našeg života, od broja koraka koje poduzimamo do financijske povijesti, su podaci. Ukupna količina podataka se stoga drastično povećava iz godine u godinu. S pojavom ogromnih količina podataka dolazi do potrebe za novijom tehnologijom, koja može obraditi velike podatke.

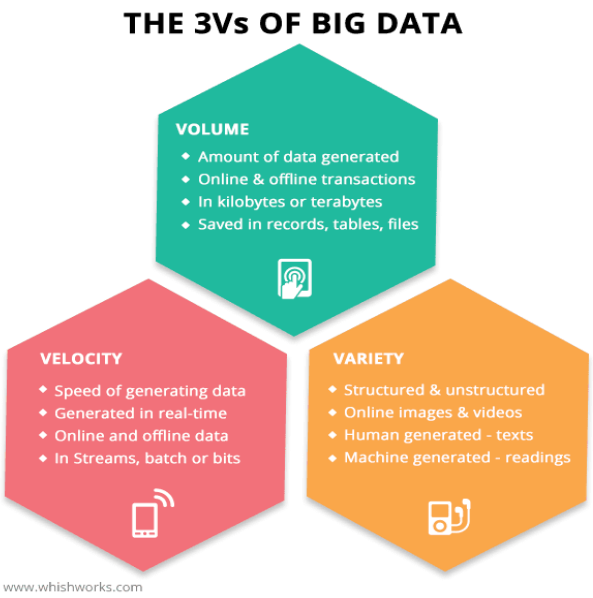
Glavni cilj ovoga rada je objasniti fenomen „big data“, definirati pojam i karakteristike, dati uvid u način rada i pojasniti ulogu digitalnih arhiva, polja koje je usko povezano s razvojem i korištenjem tehnologije velikih podataka. U prvom dijelu rada definiran je pojam „big data“ te njegove glavne karakteristike. Što velike količine podataka čini velikima odnosno bitnima? Zatim, rad se osvrće na različite tipove podataka. „Big data“ se bavi obradom, analizom i pohranom strukturiranih, nestrukturiranih i polustrukturiranih podataka, što je odvaja od tradicionalnih alata kojima su zadaća većinom bili strukturirani podaci. Rad se potom fokusira na vjerojatno najvažnije poglavlje velikih podataka, „big data“ analitiku. Radi se o principu korištenja tehnologije i načinima na koji „big data“ analitika pruža dublje razumijevanje podataka tvrtkama, organizacijama i drugim korisnicima, od razumijevanja tržišnih uvjeta, do kupovnog ponašanja kupaca, popularnosti proizvoda itd. Pružen je uvid u tipove analitike te su dani primjeri primjene odnosno uporabe određenog tipa ovisno o situaciji. Pod poglavljem analitike navedeni su najvažniji alati, njihove prednosti, povezane platforme, okviri rada, te slučajevi upotrebe u radnom okruženju. U posljednjem dijelu rada govori se o digitalnim arhivima. Opisane su definicije arhiva i što ih čini, kako arhivi prolaze proces digitalizacije, koje su prednosti digitalnog arhiviranja te razvoj arhiva u eri velikih količina podataka. Na kraju rada slijedi zaključak, popis literature te slika, te sažeci s ključnim riječima na hrvatskom i engleskom jeziku.

2. Definicija pojma „Big data“

Pojam ,,Big data“ odnosi se na veliku količinu odnosno skupove podataka koji su preveliki ili složeni da bi ih obrađivali tradicionalni alati za obradu podataka. Sam pojam je u upotrebi od 1990-ih, a s godinama dobiva na popularnosti i važnosti. „big data“ konstantno raste, od nekoliko terabajta nekad do gotovo 100 zetabajta podataka samo u 2021. godini (1 ZB = 1*10ˇ9 TB). Statistika pokazuje da mlažnjak stvara 10 terabajta podataka za vrijeme 30 minuta leta, dok društvene mreže, poput Facebook-a dnevno pohranjuju preko 500 terabajta podataka na svoje servere.[1] Takve količine podataka, koji mogu biti strukturirani, nestrukturirani ili polustrukturirani, ne samo da su ogromne veličinom, nego i eksponencijalno rastu s vremenom. Javlja se problem da današnje tvrtke ne mogu učinkovito pohranjivati i obrađivati današnje podatke koristeći se jednostavnim softverima.

Gartner 2012. daje novu definiciju opisanu preko tri ključne karakteristike pojma tzv. 3V (eng. Volumne, Velocity, Variety): „big data” su velike količine, velike brzine i/ili raznovrsne informacijske imovine koje zahtijevaju isplative, inovativne oblike obrade informacija koje omogućuju poboljšani uvid, donošenje odluka i automatizaciju procesa.[2] U proširenom modelu definicije često se ubrajaju još dvije dimenzije, a to su vrijednost (eng. Value) te istinitost (eng. Veracity) pa dobivamo takozvani 5Vs. U nastavku će biti pobliže pojašnjeno ovih pet karakteristika koje čine „big data”. U prilogu je na slici 1 prikazan 3V model.[3]

[1] Taylor D. (2009). What is Big Data? Introduction, Types, Characteristics, Examples Preuzeto s https://www.guru99.com/what-is-big-data.html (23.05.2022)
[2] Gartner Glossary. (n.d.). Preuzeto s https://www.gartner.com/en/information-technology/glossary/big-data (23.05.2022.)
[3] 3V model. Preuzeto s https://www.coforge.com/salesforce/blog/data-analytics/understanding-the-3-vs-of-big-data-volume-velocity-and-variety/ (24.05.2022.)



Slika 1: 3V model**Error! Bookmark not defined.**

2.1. Volumen

Možda je najočitija karakteristika volumen „Big data" je ogromna količina podataka koja neprestano raste. Najveći rast kroz povijest se dogodio u razdoblju od 2010. do 2012. godine, gdje se ukupna količina povećala za čak 90%.[4] Procjenjuje se da se danas pak stvara oko 2,5 kvintilijuna podataka dnevno što je otprilike povećanje od 300 puta od 2005. godine.[5] Sigurno je za reći kako svijet i dalje napreduje konstantno u elektronskoj povezanosti i napretku da će u budućnosti volumen velikih podataka narasti na brojeve koje još ni ne koristimo.

Ovakvi skupovi podataka zahtijevaju više promišljanja u svakoj fazi ciklusa obrade, analize, pohrane i razumijevanja samih podataka zbog svoje veličine. Često, budući da radni zahtjevi premašuju mogućnosti jednoga računala, to postaje izazov udruživanja, alokacije i koordinacije resursa iz skupine računala. Tvrtke prelaze s podijeljenih izvora podataka (razlomljenih metapodataka na komponente i manje podatke) na podatkovna jezera i skladišta, te na sustave upravljanja podacima. Skladišta se transformiraju s lokalnih servera u cloud platforme s vanjskim partnerima, poput Amazonovog AWS-a (Amazon Web Services). „Cloud platform odnosi se na operativni sustav i hardver poslužitelja u internetskom podatkovnom

[4] The five V's of big data. (2020). Preuzeto s https://www.bbva.com/en/five-vs-big-data/ (24.05.2022)
[5] Big Data: The 3 V's explained. (n.d.). Preuzeto s https://bigdataldn.com/news/big-data-the-3-vs-explained/ (24.05.2022.)

centru. Omogućuje koegzistiranje softverskih i hardverskih proizvoda na daljinu i u velikom broju.” [6]

2.2. Brzina

Brzina se odnosi primarno na vrijeme za koliko podaci nastaju, odnosno, u ovom slučaju, velike količine podataka nastaju u kratkom vremenskom roku. S druge strane, pojam također označava brzinu kojom se pristigli podaci obrađuju, pohranjuju i analiziraju, odnosno, koliko brzo se podaci kreću.[7] Prije informacijske revolucije podaci su pristizali na server za obradu, nakon čega se čekao rezultat analize. Ovakav način je bio održiv jer je brzina pristizanja podataka bila manja od brzine obrade. Danas velike količine podataka pristižu velikom brzinom na obradu te je bitno što prije obraditi i odvojiti podatke. Zbog kratkog roka relevantnosti podataka, tvrtke trebaju brzu obradu odnosno u gotovo realnom vremenu. Zbog učinkovitosti potrebno je podatke obrađivati odmah tijekom stvaranja, a ne samo nakon njihove pohrane.[8]

Kako je volumen podataka napravio ogroman rast, ni aspekt brzine ne zaostaje za njim. Na primjer, dovoljno je promotriti situaciju na društvenim mrežama. Informacije konstantno kruže i u nekoliko sekundi postaju viralne te s lakoćom stvaraju odjek u svijetu . Možda je to neka poslovna vijest koju će trgovci dionicama brzo interpretirati kao signal za kupovanje ili prodaju određene imovine, pritom pokrećući lavinu drugih informacija putem istih medija. Nadalje, može se raditi o kartičnom plaćanju preko interneta koje traje svega par sekundi, tijekom čega se izmjenjuju hrpe podataka o interesima kupca ili njegovim podacima među drugim stranicama, pružateljima oglasa i razni drugih. U suštini „big data“ daje moć da se podaci analiziraju ogromnom brzinom.

[6] What is a cloud platform? (n.d.). Preuzeto s: https://www.cloudbolt.io/what-is-a-cloud-platform/ (25.08.2022.)
[7] Ishwarappa, & Anuradha, J. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Computer Science, 48, str. 319-324, doi: 10.1016/j.procs.2015.04.188 (26.05.2022.)
[8] The 3V's that define big data. (n.d.). Preuzeto s http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data (26.05.2022.)

2.3. Raznolikost

Raznolikost se odnosi na razne tipove dolaznih podataka koji su dostupni. Tradicionalni podaci su strukturiranog tipa i uredno posloženi u relacijske tablice. U relacijske tablice spremaju se podaci tako da svaki dio tablice se veže za određeni skup podataka. Kako bi ovo bilo moguće potrebno je da podaci budu isključivo strukturiranog tipa. Jednostavan primjer bili bi podaci neke tvrtke o njenim kupcima. U tablici se uredno posloženo mogu pronaći podaci o imenu i prezimenu osobe koji su naravno tekstualnog oblika te datumi kupovine ili iznosi plaćanja koji bi bili u brojčanom formatu. Ovakve podatke lagano je obrađivati.

S pojavom i rastom velikih podataka, javljaju se nestrukturirani i polustrukturirani podaci u oblicima poput PDF-a, MP3, videozapisi i slično. Iako su nam jako korisni, ovakvi tipovi podataka zahtijevaju složenije analitičke vještine i dodatnu prethodnu obradu kako bi se izvuklo značenje i podržali metapodaci.[9] Dok tradicionalni alati izbjegavaju ove izazove, „big data“ tehnologija i pripadajući alati unaprijed računaju na raznovrsnost podataka.

2.4. Istinitost

Istinitost - Ovdje se radi o valjanosti i točnosti podataka. Koliko točno su upotrebljivi podaci? Nije sve od „zillion“ podataka ispravno, precizno i referentno. To je ono što je zapravo istinitost: koliko je podatak pouzdan i kvalitetan. Praktični primjer je vjerodostojnost Facebook i Twitter objava s nestandardnim akronimima ili pogreškama u pisanju. „Big data” donosi na stol mogućnost pokretanja analitike o takvoj vrsti nejasnih i nepreciznih podataka.

2.5. Vrijednost

Vrijednost – kao što ime govori, ovo je vrijednost koji podatak zapravo drži. Nije pretjerivanje ni reći da je ovo vjerojatno najvažnija „V“ ili dimenzija velikih podataka. Uostalom, glavni razlog za razvitak i iskorak prema „big data“ tehnologiji za obradu super-velikih skupova podataka je izvući neki vrijedan uvid iz njih; na kraju se ipak radi o troškovima i koristima.

[9] What is big data? (n.d.). Preuzeto s https://www.oracle.com/big-data/what-is-big-data/ (26.05.2022)

3. Vrste podataka

Veliki podaci mogu se prikupljati iz objavljenih komentara na društvenim mrežama i web stranica, dobrovoljno prikupljeni iz osobne elektronike i aplikacija ili putem upitnika kupnje proizvoda elektroničkih prijava. Prisutnost senzora i drugih ulaza u pametnim uređajima omogućuje prikupljanje podataka iz širokog spektra izvora. Postojeće skupove podataka iz vanjskih ili unutarnjih izvora mogu se klasificirati u 3 glavne skupine:

- strukturirani podaci
- nestrukturirani podaci
- polustrukturirani podaci

3.1. Strukturirani podaci

Strukturirani podaci su najlakši za korištenje. To su visoko organizirani podaci s definiranim dimenzijama i parametrima, uredno grupirani i posloženi u redove i stupce iz kojih ih je lako pronaći. Najčešće dolaze u oblicima kvantitativnih podataka poput:[10]

- dob
- kontakt
- adresa
- naplata
- troškovi
- brojevi kreditnih kartica

Kao što bi ime sugeriralo, strukturirani podaci odnosi se na podatke informacije koje imaju unaprijed definirani model podataka ili su organizirani na unaprijed određeni način. „Modeli podataka vizualni su prikazi podatkovnih elemenata poduzeća i veza među njima. Pomažući u definiranju i strukturiranju podataka, modeli podržavaju razvoj učinkovitih informacijskih sustava. Omogućuju poslovnim i tehničkim resursima da zajednički odlučuju o tome kako će se podaci pohranjivati, pristupati, dijeliti, ažurirati i koristiti u cijeloj organizaciji.“[11] Unutar tih modela podataka polja podataka koja namjeravate odabrati trebaju biti definirana i svi parametri postavljeni oko toga kako će ti podaci biti spremljeni. Na primjer, unutar tih polja

[10] Allen R. (n.d.). What are the types of big data? Preuzeto s https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/ (06.06.2022.)
[11] Erwin. (n.d.). Data Modeling: What is a data model? Preuzeto s https://www.erwin.com/products/erwin-data-modeler/ (22.08.2022)

parametri mogu biti postavljeni tako da polje telefonskog broja prihvaća samo numeričke informacije.

Strukturirani podaci su jednostavni za unos, pohranu i analizu zato što zahtijevaju vrlo malo pripreme prije procesiranja. Najizravniji i najzahvalniji su za korištenje od svih tipova podataka. Danas predstavljaju samo mali dio svih podataka u usporedbi s nestrukturiranim podacima.

3.2. Nestrukturirani podaci

S informacijskom revolucijom strukturirani podaci sve više i više ubrzano gube na važnosti jer jača tehnologija koja može procesirati nestrukturirane podatke. Može ih se jednostavno definirati kao podatke koji nemaju određenu strukturu i koje nije moguće pohraniti tradicionalnim alatima u organiziranu strukturu. Nestrukturirani podaci mogu biti generirani od ljudi ili strojno. Ovo su neki primjeri ljudski generiranih vrsta:[12]

- **E-pošta**: polja poruka e-pošte nisu strukturirana i ne mogu se raščlaniti tradicionalnim alatima za analizu
- **Tekstualni dokumenti**: obuhvaćaju word dokumente, prezentacije, blogove, log datoteke
- **Društveni mediji i web stranice**: podaci s Twittera, LinkedIna, Facebooka, YouTubea, Instagrama
- **Mobilni i komunikacijski podaci**: poruke, snimke poziva, razgovori
- **Media**: digitalne fotografije, audio i video datoteke.

Te neki od primjera nestrukturiranih podataka generiranih od strane strojeva odnosno računala:

- **Znanstveni podaci**: istraživanja nafte, Zemlje, svemira, seizmičke slike, atmosferski podaci
- **Digitalni nadzor**: izviđačke fotografije i videozapisi
- **Satelitski snimci**: vremenski podaci, vojni pokreti.

[12] Unstructured data. (n.d.). Preuzeto s https://www.mongodb.com/unstructured-data (07.06.2022.)

Naširoko se procjenjuje da danas ovaj tip obuhvaća gotovo 90% svih podataka na svijetu. Nije teško za shvatiti zašto strukturirani podaci zastupaju toliko velik dio moderne knjižnice podataka. Gotovo sve su što računala generiraju danas su restrukturirani podaci. Nitko se ne bavi prepisivanjem svakog svog telefonskog poziva, dodavanjem opisa svakoj objavljenoj fotografiji ili dodavanjem semantičkih oznaka za sve tekstualne poruke koju objave. Dok strukturirani podaci omogućuju uštedu na vremenu u analitičkom procesu, izrazito je važno dati truda i vremena da nestrukturirani podaci dobiju određenu razinu čitljivosti i vrijednosti. Kako bi dobili određenu količinu korisnih informacija, skup podataka mora biti razumljiv, a uloženi trud može biti mnogo isplativiji od jednostavnije obrade podataka.

Najteži dio kod analize nestrukturiranih podataka je usmjeriti aplikaciju da razumje koju informaciju pokušava izvući. Ovo obično znači da se nestrukturirani podaci trebaju prevesti odnosno pretvoriti u neki oblik strukturiranih podataka. Proces nije jednostavan i način na koji se izvodi varira od formata do formata i ovisi od završnog cilja analitičke obrade. Česte metode su raščlanjivanje teksta, obrada prirodnog jezika i razvoj hijerarhije sadržaja putem taksonomije.[13]

3.3. Polustrukturirani podaci

Polustrukturirani podaci[14] se odnose na podatke koji nisu formatirani na konvencionalan način. Ne slijede format tabličnog modela podataka ili relacijskih baza jer nemaju fiksnu shemu, no ipak se razlikuju od nestrukturiranih podataka jer sadrže određene organizacijske elemente. Ovi podaci sadrže karakteristike obiju prethodnih vrsta podataka pa zbog toga imaju i njihove prednosti i mane. U usporedbi sa strukturiranim podacima fleksibilniji su i jednostavniji za mjeriti, ali su i teži za rad jer je potreban dodatni trud za instruirati aplikaciju koji podatak nosi koje značenje.

HTML kodovi, grafovi i XML dokumenti su primjer ovog tipa podataka. Jezici za označavanje poput XML-a omogućavaju podacima da budu definirani po svom vlastitom sadržaju umjesto da prate određenu definiciju odnosno shemu. XML omogućuje organiziranje podataka u strukturu koja omogućuje slojevitu analizu te daje sadržajniju informaciju prikupljenu iz polustrukturiranih izvora.

---

[13] Allen R. (n.d.). What are the types of big data? Preuzeto s https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/ (06.06.2022.)

[14] What is semi-structured data? (n.d.). Preuzeto s https://www.teradata.com/Glossary/What-is-Semi-Structured-Data (09.06.2022.)

Također, moguće je koncept velike količine podataka podijeliti i iz još jedne perspektive, najviše asocirano s poslovnom inteligencijom velikih podataka i njenom prisutnošću u tvrtkama:

- **unutarnji podaci**

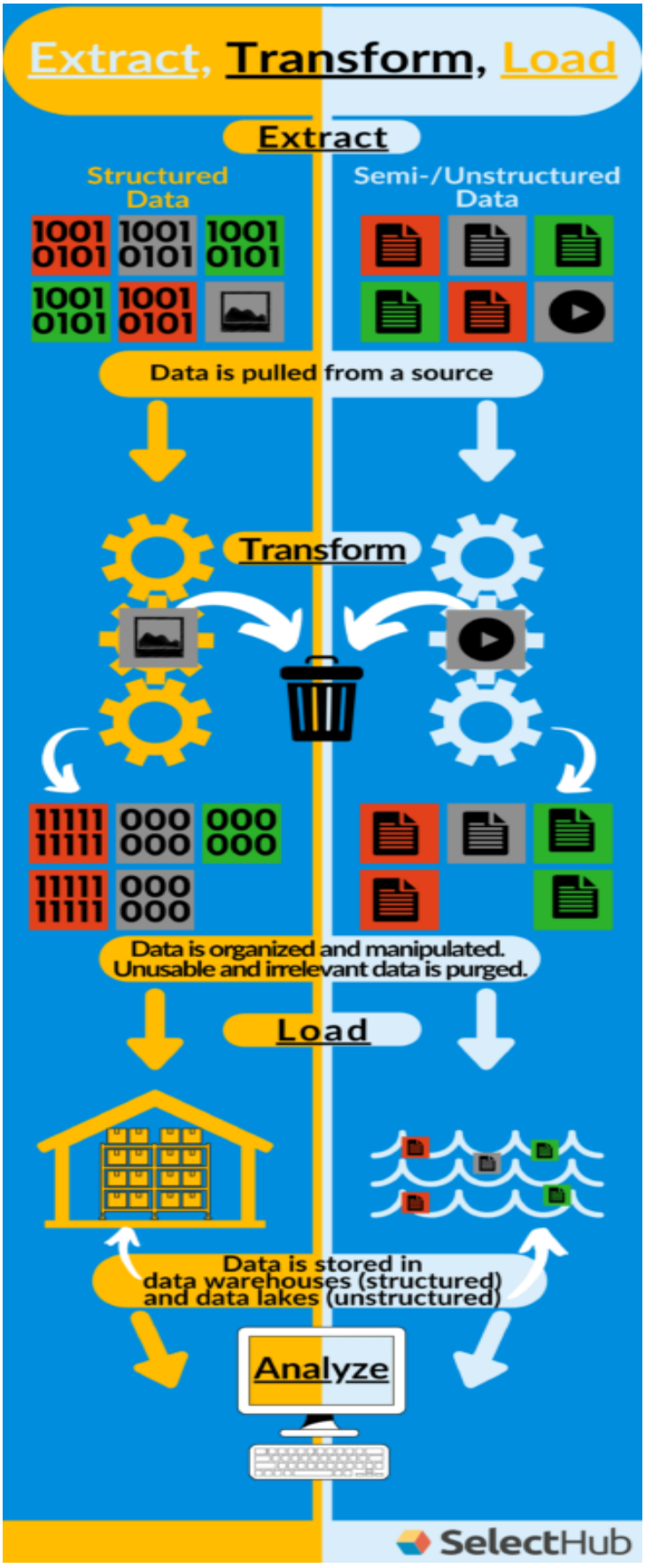
  Unutarnji ili interni podaci uračunavaju za sve što tvrtka trenutno ima ili može tome pristupiti. Ovo uključuje privatne ili vlasničke podatke koje prikupljaju i posjeduju poput: povratnih informacija korisnika, podataka o prodaji, podatci ankete zaposlenika ili kupaca, transakcijski podaci, podaci o kontroli zaliha te podaci o klijentima.

- **vanjski podaci**

  Vanjski ili eksterni podaci su beskonačan niz informacija koje postoje izvan tvrtkina poslovanja, a mogu biti privatni ili javni. Javne podatke može svatko dobiti, besplatno ili kupujući ih od treće strane, a privatne obično treba kupiti od strane tvrtke ili sporednog dobavljača. Primjer su državni podaci kao što su popis stanovništva, Facebook i Twitter podaci, Google trendovi i Google mape.

Slika 2 predstavlja pojednostavljeni shematski prikaz prijenosa podataka od ekstrakcije do analize.[15]

[15] Jezera i skladišta podataka. Preuzeto s https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/ (10.06.2022.)



Slika 2: Jezera i skladišta podataka

4. „Big data“ analitika

„Big data“ analitika odnosi se na uglavnom kompliciran proces korištenja tehnologija i tehnika pri prikupljanju, organizaciji i analizi velikih količina podataka. Glavni cilj analitike je pronaći korisne informacije koje mogu biti od koristi raznim organizacijama. Analizom podataka se traže informacije u moru podataka koje mogu dati koristan uvid kako bi učinili ispravnu poslovnu odluku, a načini na koje organizacije mogu beneficirati ovim putem su gotovo beskrajni.

Zahvaljujući tehnologiji „big data“ analitike, organizacije mogu procesirati sve tipove podataka iz višestrukih izvora. Prema Norton: „uz samu obradu skupova podatka, analitički alati i metode uključeni su u vizualizaciju, poslovno predviđanje i donošenje odluka na temelju podataka. Za razliku od tradicionalnih alata , ovi daju sirovim podacima novu dimenziju koja sadrži kontekst i značenje. Umjesto pukog spremišta pojedinačnih zapisa, „big data“ alati omogućuju organizacija sagledati širu sliku.“[16]

4.1. Kako „big data“ analitika funkcionira

„Big data“ analitika je polje koje zapošljava stručnjake iz raznih područja, analitičare, znanstvenike, modelere, statističare i druge stručnjake koje koji skupljaju, obrađuju i analiziraju strukturirane podatke kao i druge oblike podataka koje tradicionalni programi ne koriste. Proces se može opisati kroz četiri faze: [17]

1.) Prikupljanje podataka iz velikog raspona različitih izvora, kako nestrukturiranih, tako i strukturiranih podataka. Izvori podataka variraju, no neki česti izvori uključuju:
    a) podatke o internetskim klikovima
    b) zapisnike web poslužitelja
    c) cloud aplikacije
    d) mobilne aplikacije
    e) sadržaje društvenih medija
    f) tekstualne podatke elektroničke pošte i rezultata anketa

[16] Norton K. (2021). Best Big Data Tools & Software for Analytics 2022. Preuzeto s https://www.itbusinessedge.com/business-intelligence/big-data-tools/ (11.06.2022.)
[17] Chai W. Labbe M. Stedman C. (2021). Big data analytics. Preuzeto s https://www.techtarget.com/searchbusinessanalytics/definition/big-data-analytics (11.06.2022.)

g) evidencije mobilnih uređaja
h) podatke o uređaju prikupljeni senzorima priključenim na internet stvari

2.) Podaci se pripremaju i obrađuju. Nakon prikupljanja, podaci se pohranjuju u skladišta podataka ili skladišna jezera. Zadatak profesionalaca (programera, podatkovnih inženjera i arhitekata, analitičara) je da temeljito organiziraju, konfiguriraju i particioniraju podatke za analitičke upite kako bi se postigao što bolji učinak.
3.) Zatim nastupa čišćenje podataka kako bi poboljšala njihova kvaliteta. Za ovaj dio se koriste alati za skriptiranje ili softveri kako bi se podaci očistili. Pod izrazom „očistiti“ se smatra traženje bilo kakvih nepravilnosti, pogrešaka ili nedosljednosti kao što su dupliciranja ili pogreške u formatiranju te organiziranju i pospremanju podataka.
4.) Spremljeni, obrađeni i očišćeni podaci se u zadnjem koraku analiziraju analitičkim softverima. Ovo uključuje alate iz područja:
    a) dubinska analiza podataka (eng. data mining), koje pregledava skupove podataka u potrazi za obrascima i odnosima
    b) prediktivne analitike koja gradi modele za predviđanje ponašanja kupaca i drugih budućih radnji, scenarija i trendova
    c) strojnog učenja – korištenje različitih algoritme za analizu velikih količina podataka
    d) dubokog učenja – napredniji izdanak strojnog učenja
    e) rudarenja teksta i statističke analize
    f) umjetne inteligencije
    g) poslovne inteligencije
    h) vizualizacije podataka

4.2. Klasični tipovi „big data“ analitike

Podatak je informacija, a postoji određen broj načina u kakvom se obliku informacija može nalaziti ili biti prezentirana. Riječi, brojevi, slike, fotografije, videozapisi, zvukovi i razgovori su dobro poznati izvori podataka. No postoje podaci generirani iz senzora na uređajima za koje nikad ne bi rekli da su izvori informacija. Većina podataka poput onih dobivenih korištenjem senzora iz osobnih automobila, televizora ili hladnjaka, senzora za prikupljanje podataka poput temperature oceana u pet sati ujutro ili jačine udara vjetra na

mostu, ili sličnih, mogu se preobraziti u kvantitativni format koji se može dalje analizirati. Često se ti podaci kombiniraju sa strukturiranim skupovima podataka kako bi dobili na vrijednosti. Sami po sebi i dalje mogu biti zanimljivi, ali pravi uvidi dolaze kada se primjeni analitika, kombiniraju podaci i izvuče vrijednost koja seže dalje od izvornog podatka. Uobičajene vrste analiza mogu se kategorizirati po formatu podataka:[18]

- Tekstualna analiza
- Analiza govora
- Analiza slika i videa

4.2.1. Tekstualna analiza

Kao što ime sugerira, tekstualna analiza ili rudarenje teksta je proces izdvajanja informacija i uvida iz teksta. U poslovnom svijetu postoje ogromne količine tekstualnih podataka u oblicima dokumenata, izvještaja, pisane komunikacije, e-pošte, blogova, web-stranica itd. Dok ti podaci imaju smisla čovjeku, iz analitičke perspektive oni su nestrukturirani, no pružaju veliku priliku ako se na pravi način mogu iskoristiti. Na primjer, digitalizirani dokument je moguće pretraživati i tako pronaći željene informacije brzo ako se nalaze u određenom tekstu (dakako pretraga od korisnika traži da zna koju informaciju točno traži). Uglavnom, radi se o dobivanju više informacija iz teksta i pomažući da dobije na vrijednosti više od samog značenja teksta. Uključuje mnoge analitičke zadatke od kojih će važniji biti ukratko opisani u nastavku, prema Marr (2015):[19]

a) **kategorizacija teksta** – dodjeljuje dokument jednom razredu dokumenata ili više njih ovisno o predmetu ili drugim atributima kao što su vrsta dokumenta, autor, datum stvaranja itd. Kategorizacija teksta primjenjuje neku strukturu na tekst, koji se potom može koristiti za analizu ili upit.
b) **grupiranje teksta** – kako ime sugerira, grupiranje teksta omogućuje automatski grupirati ogromna spremišta teksta u smislene teme ili kategorije za brzo pronalaženje informacija ili filtriranje.
c) **ekstrakcija koncepta** – tehnika koja omogućuje izdvajanje pojmova iz teksta. Jezik može biti nejasan i značiti različite stvari ovisno o kontekstu u kojem se koristi. Ljudi

[18] Bernard Marr. (2015). Using smart big data analytics and metrics to make better decisions and improve performance. Wiley (12.06.2022.)
[19] Ibid.

mogu lako razumjeti što se podrazumijeva pod kontekstom i okolnim riječima. Računalo može učiniti slično kada ima puno podataka za obradu kako bi ujednačilo u rezultatima i smanjile greške.

d) **analiza sentimenta** – teži izvući subjektivno mišljenje ili sentiment iz teksta. Nastoji analizom dijelove teksta ili cijeli tekst kategorizirati po osjećajnoj polarnosti na pozitivno, negativno ili neutralno. Tvrtkama daje uvid o mišljenju kupaca prema njihovom proizvodu.
e) **sažimanje dokumenata** – automatski sažima dokument pomoću računalnog programa i zadržava najvažnije točke iz izvornih dokumenta. Postoje dva pristupa: ekstrakcija i apstrakcija. Ekstrakcija radi sažetak odabirom postojeći izraza i rečenica, a apstrakcija gradi unutarnju semantičku reprezentaciju, a zatim koristi tehnike generiranja prirodnog jezika za stvaranje sažetka. Apstrakcijom stvoren sažetak je bliže onome što bi čovjek mogao proizvesti.

4.2.2. Analiza govora

Govor može isto kao i tekst biti podvrgnut analizama. Analizom govora mogu se identificirati teme diskusije, ali i razotkriti emocionalni sadržaj razgovora. Korist ove tehnologije može se vidjeti na primjeru poduzeća koja ju koriste kako bi došli do vrijednih informacija o proizvodima, kupcima, operativnim problemima te stanju korisničke službe. Iz takvih razloga razgovori kupaca s korisničkom službom neke tvrtke često sadrže upozorenje da mogu biti snimani. Ta snimka će biti pohranjena i analizirana da bi se dobio uvid kako se korisnici osjećaju, kakav stav imaju prema tvrtki i koje su najčešće teme i problemi s kojima se suočavaju. Svi ti faktori zajedno imaju veliku poslovnu stratešku vrijednost. Alati za analitiku govora mogu se naći u komercijalnim softverima za prepoznavanje govora, diktafonima ili aplikacijama na pametnim telefonima. Omogućavaju nam mogućnost pretvorbe govora u pisanu riječ svakodnevno kada koristimo na tražilicama alate za pretragu govorom ili sustav Siri na iPhone mobitelu, ili obrnut proces, gdje sustav iščitava dobivenu pisanu poruku s uređaja. Također, zbog mogućnosti da prepozna emocije poput stresa, straha, sreće ili tuge u nečijem glasu, te laže li neka osoba ili ne, analiza govora postaje područje koje ima sve veću ulogu i u rješavanju slučajeva kriminala i prevenciji prevara.

4.2.3. Analiza slika i videa

U prošlim vremenima većina video podataka koji su prikupljani bili su sigurnosni podaci. Zbog ograničenih mogućnosti pohrane videozapisa, sigurnosni sistemi bi snimke kamera vrtjeli u petlji, gdje bi kamera snimala određeni period bez prestanka i onda počela ponovno snimati preko starijeg materijala koji se ne bi sačuvao. To se mijenja s razvojem tehnologije i procesorske snage računala. Nadalje, većina prave analitike videozapisa ili slika se odvijala putem oznaka koje ih opisuju. Tako da kada bi prije pretražili net ili YouTube za neki pojam, sustav bi pokušao identificirati videozapise preko oznake za pojam koji smo tražili, nadajući se da će se informacija nalaziti u naslovu ili opisu dokumenta. Danas analitički alati koriste algoritme koji prolaze temeljito kroz videozapise, označuju informacije zasebno i identificiraju uzorke među njima putem drugih alata. Neke od bitnih primjena video analize sadržaja su:[20]

- prepoznavanje lica – računalna aplikacija koja može automatski identificirati i provjeriti osobu iz digitalne slike ili videa.
- analiza ponašanja – služi za mjerenje i praćenje ponašanje osobe
- svjesnost situacije –upotreba ove analitike je u kompliciranim, fluidnim situacijama za donošenje odluka poput kontrole leta, navigacije brodova, hitni službi itd.

4.3. „Big data“ analitika prema slučaju uporabe

Osim podjele po formatu i tipu podataka, različite vrste „big data“ analiza postoje ovisno o slučaju upotrebne odnosno razlogu korištenja. U nastavku će biti ukratko opisana četiri tipa te odgovarajući primjer za svakoga.[21]

a) **deskriptivna analitika** – sažima podatke iz prošlosti u oblik koji ljudi mogu lako pročitati. Pomaže u stvaranju izvješća primanja i dobiti neke tvrtke, također pomaže u tabeliranju metrike društvenih medija. Primjer: Dow Chemical Company je analizirala svoje prošle podatke kako bi povećala iskorištenost svog laboratorijskog i uredskog prostora. Ovo je rezultiralo konsolidacijom nedovoljno iskorištenog prostora koja je uštedjela tvrtki 4 milijuna dolara godišnje.
b) **dijagnostička analitika** – koristi se kako bi se razumjelo gdje je izvor određenog problema. Ovo obuhvaća tehnike poput „drill-down“ dubinske analize i oporavka

[20] Ibid, str. 13.
[21] What is Big Data Analytics and Why It is Important? (2022). Simplilearn. Preuzeto s https://www.simplilearn.com/what-is-big-data-analytics-article (14.06.2022.)

podataka. Organizacije koriste dijagnostičku analitiku jer pruža dubinski uvid u određeni problem. Primjer: Online trgovina je ovim putem ustanovila da im prodaja pada, ali korisnici dodaju proizvode u košaricu više. Zaključak je da možda treba popraviti stanje web-stranice, dodati više platnih opcija ili je cijena dostave prevelika.

c) **prediktivna analitika** – grana analize koja gleda u prošle i sadašnje podatke kako bi napravila pretpostavke za budućnost. Koristi se rudarenjem podataka, umjetnom inteligencijom i strojnim učenjem. Primjer: PayPal prediktivnom analitikom koriste sve podatke iz prošlosti o transakcijama i ponašanju korisnika kako bi poboljšali svoj algoritam koji štiti korisnike i pokušava predvidjeti prijevarne aktivnosti.

d) **preskriptivna analitika** – „Preskriptivna analitika je proces korištenja podataka za određivanje optimalnog tijeka djelovanja.“[22] Pripisuje rješenje određenom problemu. Radi se mješavini deskriptivne i prediktivne analitike a većinu vremena oslanja se na umjetnu inteligenciju i strojno učenje. Primjer: zrakoplovne tvrtke maksimiziraju svoje profite tako što izgrade algoritam koji će automatski podesiti cijene letova temeljeno na čimbenicima poput potražnje kupaca, vrijeme, odredište, cijene nafte itd.

4.4. Alati za „big data“ analize

Infrastrukturne tehnologije su sama srž „big data“ ekosustava. Infrastruktura velikih podataka uključuje alate koji prikupljaju podatke, softverske sustave i fizička skladišta koji ih pohranjuju, mrežu koja ih prenosi, aplikacijska okruženja koja ugošćuju analitičke alate koji ih analiziraju te arhivsku infrastrukturu koja pruža podršku. Organizacije su se desetljećima oslanjale na relacijske baze podataka, klasične skupine redova i stupaca, kako bi obrađivale strukturirane podatke. S rastom volumena, brzine i raznolikosti nestrukturiranih podataka, došlo je do promjene i inovacija u tehnologiji, što je omogućilo napredak u ovom polju. Pojavile su se nove tehnologije sposobne suočiti se s velikim količinama složenih podataka. U nastavku će biti spomenuti i okratko opisani neki od danas najkorištenijih i najutjecajnijih alata.

- NoSQL – stoji za „ne samo SQL“ (eng. Not only SQL). Ovi alati obrađuju velike volumene različito strukturiranih podataka. SQL se desetljećima razvijao na relacijskim bazama podataka dok se NoSQL počinje spominjati i koristi tek u 21. stoljeću za potrebe web 2.0 kompanija. NoSQL baze podataka često se koriste u web aplikacijama

[22] Cote C. (2021). What is prescriptive analysis? Harvard buisness school online. Preuzeto s https://online.hbs.edu/blog/post/prescriptive-analytics (23.08.2022.)

i to u stvarnom vremenu. Prema McNulty-Holmes „ ideja da su SQL i NoSQL u izravnoj suprotnosti i međusobna konkurencija nije točna, a pokazatelj toga je i činjenica da se mnoge tvrtke odlučuju istodobno koristiti ove dvije vrste alata. Ne postoji pristup „jedan sustav za sve“, a izbor prave tehnologije ovisi o slučaju korištenja.“[23]

- Apache Hadoop je projekt Apache software foundation-a, jedne od najstarijih i utjecajnih organizacija na ovom području. Hadoop je analitički otvoreni softver, prvi put je dan na preuzimanje u 2006. godini i još uvijek je jedan od najpopularnijih alata. Sastoji se od mnogo komponenti, no mogu se izdvojiti četiri ključne: [24]
- Hadoop Common – skup knjižnica i uslužnih programa koji koriste Hadoop moduli
    1.) HDFS – zadani sloj za pohranu za Hadoop
    2.) MapReduce – izvršava širok raspon analitički funkcija paralelno analizirajući skupove podataka prije „smanjivanja“ rezultata
    3.) YARN – klaster upravljački sloj Hadoop-a

  Hadoop je dizajniran s temeljnim razumijevanjem da su kvarovi hardvera neizbježni, tako da sustav treba biti spreman za otkrivanje i rješavanje problema na sloju aplikacije. Iako Hadoop nudi visoku dostupnost i iznimne mogućnosti obrade, ima i nedostatke poput nepodržavanja obrade u stvarnom vremenu ili izračuna memorije, a obje su presudne za učinkovitu analizu podataka. Organizacija Apache je također izdala druge alate koje nadopunjuju Hadoop i kompenziraju njegove nedostatke. Alati poput Apache Spark i Apache Storm su među najkvalitetnijim analitičkim alatima za obradu podatka danas.

- MongoDB –open-source NoSQL baza podataka koja operira kao napredna alternativa modernim bazama podataka. To je dokumentno orijentirana baza podataka koja se koristi za pohranjivanje velikih podataka. Umjesto redaka i stupaca koji se koriste tradicionalno, MongoDB koristi dokumente i zbirke.[25] Ovaj alat je idealan za tvrtke koje trebaju donositi brze poslovne odluke i raditi s podacima u stvarnom vremenu. Prednosti alata su što je jako fleksibilan i lako se može adaptirati jer sprema podatke u dokumente. Podržava razne tipove pretrage, sva polja MongoDB dokumenta mogu se

[23] McNulty-Holmes, E. (2014). Understanding Big Data: A begginers guide to Data science & the business applications. Eileen McNulty-Holmes (15.06.2022.)
[24] Norton K. (2021). Best Big Data Tools & Software for Analytics 2022. Preuzeto s https://www.itbusinessedge.com/business-intelligence/big-data-tools/ (11.06.2022.)
[25] Sharma R. (2021). Big data tools. Preuzeto s https://www.upgrad.com/blog/big-data-tools/ (16.06.2022.)

indeksirati radi poboljšanja kvalitete pretraživanja. Dopušta pohranu podataka bilo kojeg tipa poput cijelih brojeva, nizova, booleana itd. Kako ova tehnologija koristi dinamičnu shemu rada, moguće je pohranjivati i pripremiti podatke brzo. Tvrtke koje traže ovakav alat ga vole jer im radi i uštedu, vrijeme je novac.

- Cassandra – distribuirani sustav upravljanja bazama podataka koji se preferirano koristi za obradu skupova strukturiranih podataka. Razvijen je od Facebooka prvotno kao rješenje za NoSQL,a trenutno ovaj sustav koriste korporacijski divovi poput Netflixa, Twittera te Cisca.[26] Pogodnosti ovog alata su pružanje jednostavnog jezika upita, što korisnicima olakšava prijelaz s relacijskih baza podataka na Cassandru. Arhitektura sustava omogućava čitanje i pisanje podataka na bilo kojem čvoru. Ti podaci se repliciraju na različitim čvorovima tako da iako jedan čvor ne radi, podaci će biti dostupni za korištenje na drugim čvorovima. Podaci se također mogu replicirati u više podatkovnih centara. Ako se izgube ili oštete na jednom centru, mogu se dohvatiti s drugoga. Ima ugrađene sigurnosne značajke, kao što su mehanizam vraćanja i sigurnosna kopija podataka. Cassandra se za sada naširoko koristi u aplikacijama internata stvari (IoT) u stvarnom svijetu gdje ogromni tokovi podataka dolaze s uređaja i senzora te također za analizu društvenih medija.

[26] Ibid, str. 17.

5. Digitalno arhiviranje

Danas je pojam „digitalnog“ način života i ljudi se svakodnevno služe s digitalizacijom. Naime, tako je i s elektroničkim dokumentima te elektroničkim zapisima, bilo da je riječ o privatnom radu određene osobe ili poslovnim procesima. „Životni ciklus elektroničkih dokumenata, od njihova nastanka pa sve do odlaganja u digitalni arhiv, odvija se u određenim fazama kojima je potrebno upravljati kao procesima.“[27] Digitalni arhiv nikako ne može spontano nastati ili biti neplanski formiran. „Proces uspostave digitalnog arhiva uvijek prolazi kroz sve faze razvoja informacijskog sustava:

- planiranje,
- analizu,
- oblikovanje,
- izradu,
- uvođenje u rad i
- održavanje.“[28]

Postavlja se pitanje, što je zapravo definicija digitalnog arhiva? „Dakle, može se reći kako je digitalni arhiv po namjeni sličan fizičkom arhivu, a povijesni dokumenti i objekti koji pružaju dokaze o prošlosti su digitalizirani.“[29] Dokumenti se mogu arhivirati skeniranjem ili fotografiranjem, osim ako je dokument već napravljen na digitalni način i stavljeni na raspolaganje na internetu. Digitalni arhivi se obično stvaraju s ciljem očuvanja povijesnih predmeta i stavljanja na raspolaganje istraživačima. Drugo važno pitanje oko digitalnih arhiva jest tko stvara digitalne arhive? Mnogi digitalni arhivi spadaju u jednu od tri neslužbene skupine, a to su oni koje osiguravaju fizički arhivi, knjižnice, sveučilišta, muzeji, državne agencije ili druge povijesne odnosno kulturne udruge.

Digitalni arhivi mogu biti i privatni kako je već u početku ovog poglavlja navedeno. Naime, neke digitalne arhive mogu kreirati pojedinci, male grupe ili neprofitne organizacije koje dijele interese kao što su primjerice članovi obitelji koji digitaliziraju zanimljive povijesne materijale iz svoje obitelji. Razlika između ovog privatnog arhiviranja i već gore navedenog

[27] Rajh A., Stančić H. (2010). Planiranje, izgradnja i uspostava digitalnog arhiva. Arh. vjesn., god. 53, str. 41-62. Preuzeto s https://hrcak.srce.hr/62414 (17.06.2022.)
[28] Ibid.
[29] Jaillant L. (2022). How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions. Arch Sci 22, str. 417–436. doi: 10.1007/s10502-022-09390-7 (17.06.2022.)

jest u kvaliteti i autoritetu određenih zbirki jer privatne zbirke nisu uvijek tako pouzdane kao zbirke koje stvaraju institucije.[30]

„Sva digitalna arhivska građa je vrijedna.“[31] Dakle, arhivsko gradivo, tj. određene zbirke sastavlja izdavač ili neka druga profitna tvrtka koja naplaćuje naknadu za kupnju ili pretplatu na zbirku. „Neke zbirke mogu se prodavati pojedincima, osobito manje zbirke ili one s određenim genealoškim fokusom, ali veće i skuplje zbirke često se prodaju samo knjižnicama.“[32] Važno je za naglasiti da je arhivsko gradivo većinom besplatno za pregledavanje.

„Potrebno je promijeniti paradigmu čuvanja i shvatiti da će elektroničke zapise tijekom čuvanja biti potrebno migrirati na nove medije zbog potencijalne zastarjelosti tehnologije.“[33] „Značaj digitalnog arhiva, treba promatrati kao početnu točku procesa. To znači da digitalni arhiv ne egzistira sam, već se uvijek nalazi u okviru neke institucije i uvijek je dio nekog procesa dugoročnog ili trajnog čuvanja elektroničkih zapisa.“[34]

Koja je prednost digitalnih arhiva? Naime, u današnje vrijeme se svijet trudi biti ekološki prihvatljiv. Značajan korak koji ljudi poduzimaju uključuje ne korištenje papira i stavljanje većine svojih procesa na internet. Digitalno arhiviranje jedna je stvar na koju se tvrtke kreću i koja služi objema gore spomenutim svrhama.[35]**Error! Bookmark not defined.** Digitalno arhiviranje je proces pretvaranja tiskanog primjerka u digitalni oblik ima nekoliko pozitivnih primjera za što je tako popularan.

Prva pozitivna stavka je ta da se podaci ne mogu izgubiti. Lako je izgubiti podatke kada su u opipljivom obliku, kada postoji puno papirologije koja stvara brigu, a datoteka može biti izgubljena ako se ne čuva na sigurnom. Također, pronalaženje informacija iz nekoliko datoteka i mapa može biti teško i dugotrajno.[36] To nije slučaj s digitalnim arhiviranjem jer je sve

[30] Murphy H. (2022). Understanding the Value of Digital Archival Collections to Faculty at Maynooth University Library. Academic librarianship, 41, str. 423-439. doi: 10.1080/13614533.2021.1976233 (18.06.2022.)
[31] Ađulović S. (2019). Arhivi i digitalizacija – od mikrofilma do digitalnog arhiva. @rhivi, Vol. 5, str. 26-28. Preuzeto s https://hrcak.srce.hr/236310 (18.06.2022.)
[32] Jaillant L. (2022). How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions. Arch Sci 22, str. 417–436. doi: 10.1007/s10502-022-09390-7 (17.06.2022.)
[33] Rajh A., Stančić H. (2010). Planiranje, izgradnja i uspostava digitalnog arhiva. Arh. vjesn., god. 53, str. 41-62. Preuzeto s https://hrcak.srce.hr/62414 (17.06.2022.)
[34] Ibid.
[35] Murphy H. (2022). Understanding the Value of Digital Archival Collections to Faculty at Maynooth University Library. Academic librarianship, 41, str. 423-439. doi: 10.1080/13614533.2021.1976233 (18.06.2022.)
[36] Jaillant L. (2022). How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions. Arch Sci 22, str. 417–436. doi: 10.1007/s10502-022-09390-7 (17.06.2022.)

dostupno u nekoliko klikova, a ako se dobro sortiraju svi podaci, sve se može imati u odgovarajućim mapama.

Druga pozitivna stavka je ekologija, odnosno da je digitalno arhiviranje ekološki prihvatljivo. „Uobičajena metoda spremanja podataka troši mnogo resursa. Kao prvo, mora se potrošiti puno novca na kupnju papira. Upotreba papira ne smatra se ekološki prihvatljivim izborom i korištenje papira svakog mjeseca nije ništa drugo nego rasipanje resursa.“[37] S te ekološke strane, nezamislivo je koliko se stabala posiječe samo za određenu količinu papira. Dakle, kada se prijeđe na digitalno arhiviranje, čini se korak prema ekološkoj prihvatljivosti. Digitalno arhiviranje ne uključuje korištenje papira i drugih materijala koji mogu uništiti okoliš.

Treća pozitivnost digitalnog arhiviranje je potpuna sigurnost podataka. Naime, puno povjerljivih podataka koji se ne trebaju dijeliti s okolinom treba dobro čuvati. Opipljive datoteke i dokumenti uvijek su u opasnosti da bi mogli doći u pogrešne ruke. Digitalno arhiviranje štiti podatke. Dokumenti se mogu zaštititi i ograničiti im pristup sigurnosnim mjerama koje su u digitalnom svijetu normalne.[38]

Četvrta pozitivna strana digitalnosti kod arhiviranje je brži pristup podacima. Digitalno arhiviranje je ušteda vremena. Svi dokumenti i podaci mogu biti dostupni u nekoliko klikova za razliku od korištenja opipljivih dokumenata.[39]**Error! Bookmark not defined.**

Vrlo važna i posljednja pozitivnost u nizu pozitivnih stavki je šansa za povrat digitalno arhivirane građe. Kod fizičke metode spremanja i očuvanja podataka je problem vratiti dokument ako je uništen, to je nemoguće ako ne postoji kopija.[40] Kod digitalnog arhiviranja još postoji prilika za vraćanjem podataka.

„Ipak, digitalni arhiv u odnosu na druge informacijske sustave ima dodatne i složenije zahtjeve. On mora ne samo čuvati elektroničke zapise kroz dulje vrijeme, već ponekad i trajno. U slučaju elektroničkih zapisa to je znatno teže učiniti nego kod klasičnih, analognih zapisa.“[41] „Pritom ne treba smetnuti s uma da će za neke sačuvane zapise biti potrebno očuvati i njihovu autentičnost kroz sve potrebne, tijekom duljeg vremena potencijalno višestruke, prikladne

[37] Murphy H. (2022). Understanding the Value of Digital Archival Collections to Faculty at Maynooth University Library. Academic librarianship, 41, str. 423-439. doi: 10.1080/13614533.2021.1976233 (18.06.2022.)
[38] Ibid.
[39] Ibid.
[40] Ibid.
[41] Rajh A., Stančić H. (2010). Planiranje, izgradnja i uspostava digitalnog arhiva. Arh. vjesn., god. 53, str. 41-62. Preuzeto s https://hrcak.srce.hr/62414 (17.06.2022.)

postupke čuvanja.“[42] Dakle, digitalni arhivi nisu neka mjesta na koja se može odložiti zapise i više na njih ne obraćati pažnju sve dok ponovno ne budu potrebni, već su to sustavi o kojima je potrebno proaktivno brinuti. S tim rečenim, evidentno je da postoje posebni izazovi digitalnog arhiviranja, a potrebno je spomenuti:[43]

- Zastarjelost – kao i analogni, digitalni arhivi mogu zastarjeti i postati nečitljivi. Ova opasnost postoji zbog konstantnog tehnološkog napretka. Digitalni podaci moraju ostati interoperabilni i obradivi.
- Vrijeme i troškovi održavanja – arhivi se ne čuvaju samo na duge periode, već i vječno. Iako digitalizirani, računalna oprema nije besplatna, a optimizacija uvijek postoji i dugoročne troškove je teško izračunati.
- Zaštita i osiguranje objekta – u kontrastu prema analognim arhivima, kod digitalnih arhiva nema razloga za strah od gubitka kvalitete podataka s vremenom. No potrebno je napraviti sigurnosne kopije svakako, a postoji problem dugoročne zaštite protiv neautoriziranog pristupa od strane hakera ili virusa, stoga sustave treba nanovo ažurirati i unapređivati.
- Efikasnost – „Ako su javni sadržaji arhiva također dostupni na Internetu, istraživanje je olakšano korisnicima.“[44] Moguće je pretraživati sadržaje putem mreže što stvara uštedu na novcu i vremenu te rasterećuje arhiv od posjetitelja. No nisu svi analogni sadržaji pretvoreni u digitalne te još postoje ogromne količine podataka koje čekaju digitalizaciju.
- Prostorna zahtjevnost – „u usporedbi s tradicionalnim arhivima, digitalni arhivi imaju znatno smanjenu potrebu za prostorom pohrane kako su informacije prilično kompresirane.“[45] Ipak, malo je vjerojatno da će se nekonvencionalni arhivi u potpunosti odreći analognih originala.

5.1. Digitalni arhivi u „big data“ eri

Pojava računalne tehnologije, posebice internetske tehnologije, pomiče uspon digitalizacije i donosi revoluciju digitalne memorije za pohranu. Strategija velikih podataka

[42] Ibid, str. 22.
[43] Ghosh A. (2021). What is a digital archive? Preuzeto s https://thecustomizewindows.com/2021/10/what-is-a-digital-archive/ (20.06.2022.)
[44] Ibid.
[45] Ibid.

ima dubok utjecaj na koncept, tehnologiju i način arhivskog rada - te promiče da je potrebno usvojiti velike podatke, inteligentno upravljanje i druge tehnologije za poboljšati informatizaciju arhiva, duboki razvoj i razinu usluge arhivskih informacijskih resursa.[46]

Digitalizacija je vrlo čest pojam u digitalnom dobu. „Digitalizacija je postupak kojim se analogni procesi i fizički objekti pretvaraju u digitalni format. Drugim riječima, digitalizacija je postupak kojim se određene operacije mogu početi provoditi putem digitalnih medija, poput računala ili pametnih telefona, obično uz pomoć internetske veze.“[47] „Digitalizacija promovira društveni digital (odvijanje socijalnih radnji u digitalnom okruženju), dakako, digitalizacija donosi analogni svijet u digitalno okruženje, što omogućuje ljudskom društvu pohranjivanje više informacija i bržu obradu.“[48]

„Dugo vremena, arhivski odjeli su razvijali razne tradicionalne modele usluga, poput pristupa datotekama, kompilacije datoteka i istraživanja, sortiranje spisa, certifikacije, konzultacije, reference i nadalje ali ove usluge teško zadovoljavaju potrebe korisnika arhiva danas, obzirom na praktičnost, učinkovitost, znanje, personaliziranu uslugu i mrežnu povezanost u novoj eri."[49] U eri velikih količina podataka neki znanstvenici smatraju da je neophodno koristi tehnologiju velikih podataka za analizu strukturiranih, nestrukturiranih i polustrukturiranih podataka poput identiteta korisnika, posuđivanja sadržaja, ponašanja pohrane, metoda pretraživanja, zapisa riječi i dijela, rudarenja i predviđanja implicitnih zahtjeva korisnika, što omogućuje nadogradnju razine usluge, ostvariti humanizaciju i znanje arhivske usluge te promijeniti način usluge orijentiran na potražnju u servisni način orijentiran na korisnike.

Razvoj arhivskih resursa temelji se na procesu obrade raznih nositelja i oblika arhiva i arhivskih kolekcija radi zadovoljavanja različitih potreba u cilju formiranja bolje usluge. Temeljna svrha arhivskog rada je duboko istražiti vrijedne informacije sadržane u arhivskim izvorima, pronaći i nabaviti više sustavno ili specifično vrijednosno znanje i mudrost, te učinkovito pružiti korisnicima arhiva specifične potrebe u različitim područjima društva. Tradicionalne arhivske usluge u velikoj mjeri ovise o dubini indeksiranja i kompilaciji alta za pronalaženje kao što je katalog dokumenata, katalog datoteka, tematski katalog, vodič itd. „Uz

[46] Akella J. (2016). The value promotion of arhives management in the era of big data. Revista Argentina de Clínica Psicológica 2021, Vol. XXX, str. 250-258 doi: 10.24205/03276716.2020.4023 (21.06.2022.)
[47] Digitalizacija. (n.d.). Preuzeto s https://hr.economy-pedia.com/11041145-digitization (24.08.2022.)
[48] Odour MD. (1990). Resarch on knowledge mining technology in big data anylsis of digital archives Revista Argentina de Clínica Psicológica 2021, Vol. XXX, str. 250-258 doi: 10.24205/03276716.2020.4023 (21.06.2022.)
[49] Ibid.

razvoj digitalnih arhiva, indeksiranje digitalnih arhivskih resursa temeljenih na metapodacima je postala zrela i uhodana tehnologija, što je od velikog značaja za opis, pronalaženje i dugotrajno čuvanje arhivskih resursa.“[50]

[50] GUO Wei-ya. (2021). Prospect of Archive Dataization in Big Data Age: Significance and Dilemma. Revista Argentina de Clínica Psicológica 2021, Vol. XXX, str. 250-258 doi: 10.24205/03276716.2020.4023 (21.06.2022.)

6. Zaključak

U ovom radu prikazane su razne uloge i implementacije koncepta velike količine podataka, prvo teoretski, a kasnije putem primjera i slučajeva korištenja. Raspravlja se o velikim podacima, kako i gdje se prikupljaju te kakvi se analitički alati koriste za obradu, analizu i pohranu tih podataka. „Big data“ analitika pruža duboki uvid u poslovanja različitih organizacija. Kada je ispravno primijenjena, omogućava mnoge pogodnosti poput optimizacije resursa, boljeg upravljanja imovinom, skraćivanja vremena reakcije pri otklanjanju problema, poboljšanja korisničke usluge itd. Analitičkih alata ima mnogo, s konstantnim razvojem tehnologije ne manjka rješenja za razne potrebe upravljanja velikim podacima. Ipak, brzina kojom se novi zahtjevi pojavljuju traži određenu dozu opreza pri izboru alata i platformi za obradu, analizu i pohranu podataka. Bitno je pažljivo izabrati tehnologiju prema potrebama kako bi se optimizirala usluga, te gledati na budućnost kako će se određena platforma s vremenom nositi s operativnim zadacima. Držanje koraka s tehnologijom je neophodno, no sigurno je da su veliki podaci promijenili igru te će se ovo područje nastaviti razvijati u budućnosti. Količina podataka već je masivna, ali je za očekivati da će rasti eksponencijalno u budućnosti s obzirom na svakodnevni razvoj nove tehnologije. Uostalom, većina svih podataka u pohrani je nastala tijekom zadnjih par godina. „Big data“ i poslovna analitika su danas „mainstream“ tehnologija, a s automatizacijom i umjetnom inteligencijom predstavljaju temelj na kojem se gradi proces digitalne transformacije. Arhivistika je jedna od znanstvenih grana koja se susrela s procesom velike preobrazbe i adaptacije na novu eru arhiviranja. „Big data“ paradigma promiče da je potrebno usvojiti velike podatke unutar arhivistike, provesti digitalizaciju arhivskog sadržaja i arhivskih usluga kako bi se poboljšala informatizacija arhiva, duboki razvoj i razina usluge informacijskih resursa.

Popis literature

1. Ađulović S. (2019). Arhivi i digitalizacija – od mikrofilma do digitalnog arhiva. @rhivi, Vol. No. 5, str. 26-28. Preuzeto s https://hrcak.srce.hr/236310 (18.06.2022.)
2. Akella J. (2016). The value promotion of arhives management in the era of big data. Revista Argentina de Clínica Psicológica 2021, Vol. XXX, str. 250-258 doi: 10.24205/03276716.2020.4023 (21.06.2022.)
3. Allen R. (n.d.). What are the types of big data? Preuzeto s https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/ (06.06.2022.)
4. Bernard Marr. (2015). Using smart big data analytics and metrics to make better decisions and improve performance. Wiley (12.06.2022.)
5. Big Data: The 3 V's explained. (n.d.). Preuzeto s https://bigdataldn.com/news/big-data-the-3-vs-explained/ (24.05.2022.)
6. Chai W. Labbe M. Stedman C. (2021). Big data analytics. Preuzeto s https://www.techtarget.com/searchbusinessanalytics/definition/big-data-analytics (11.06.2022.)
7. Cote C. (2021). What is prescriptive analysis? Harvard buisness school online. Preuzeto s https://online.hbs.edu/blog/post/prescriptive-analytics (23.08.2022.)
8. Digitalizacija. (n.d.). Preuzeto s https://hr.economy-pedia.com/11041145-digitization (24.08.2022.)
9. Erwin. (n.d.). Data Modeling: What is a data model? Preuzeto s https://www.erwin.com/products/erwin-data-modeler/ (22.08.2022)
10. Gartner Glossary. (n.d.). Preuzeto s https://www.gartner.com/en/information-technology/glossary/big-data (23.05.2022)
11. Ghosh A. (2021). What is a digital archive? Preuzeto s https://thecustomizewindows.com/2021/10/what-is-a-digital-archive/ (20.06.2022.)
12. GUO Wei-ya. (2021). Prospect of Archive Dataization in Big Data Age: Significance and Dilemma. Revista Argentina de Clínica Psicológica 2021, Vol. XXX, str. 250-258 doi: 10.24205/03276716.2020.4023 (21.06.2022.)

13. Ishwarappa, & Anuradha, J. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Computer Science, 48, str. 319-324 doi: 10.1016/j.procs.2015.04.188 (26.05.2022.)

14. Jaillant L. (2022). How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions. Arch. Sci.22, str. 417–436. doi: 10.1007/s10502-022-09390-7 (17.06.2022.)

15. McNulty-Holmes, E. (2014). Understanding Big Data: A begginers guide to Data science & the business applications. Eileen McNulty-Holmes (15.06.2022.)

16. Murphy H. (2022). Understanding the Value of Digital Archival Collections to Faculty at Maynooth University Library. Academic librarianship, 41, str. 423-439. doi: 10.1080/13614533.2021.1976233 (18.06.2022.)

17. Norton K. (2021). Best Big Data Tools & Software for Analytics 2022. Preuzeto s https://www.itbusinessedge.com/business-intelligence/big-data-tools/ (11.06.2022.)

18. Odour MD. (1990). Resarch on knowledge mining technology in big data anylsis of digital archives Revista Argentina de Clínica Psicológica 2021, Vol. XXX, str. 250-258 doi: 10.24205/03276716.2020.4023 (21.06.2022.)

19. Rajh A., Stančić H. (2010). Planiranje, izgradnja i uspostava digitalnog arhiva. Arh. vjesn., god. 53 , str. 41-62 Preuzeto s https://hrcak.srce.hr/62414 (17.06.2022.)

20. Sharma R. (2021). Big data tools. Preuzeto s https://www.upgrad.com/blog/big-data-tools/ (16.06.2022.)

21. Taylor D. (2009). What is Big Data? Introduction, Types, Characteristics, Examples

22. Preuzeto s https://www.guru99.com/what-is-big-data.html (23.05.2022)

23. The 3V's that define big data. (n.d.). Preuzeto s http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data (26.05.2022.)

24. The five V's of big data. (2020). Preuzeto s https://www.bbva.com/en/five-vs-big-data/ (24.05.2022)

25. Unstructured data. (n.d.). Preuzeto s https://www.mongodb.com/unstructured-data (07.06.2022.)

26. What is a cloud platform? (n.d.). Preuzeto s: https://www.cloudbolt.io/what-is-a-cloud-platform/ (25.08.2022.)

27. What is Big Data Analytics and Why It is Important? (2022). Simplilearn. Preuzeto s https://www.simplilearn.com/what-is-big-data-analytics-article (14.06.2022.)

28. What is big data? (n.d.). Preuzeto s https://www.oracle.com/big-data/what-is-big-data/ (26.05.2022)

29. What is semi-structured data? (n.d.). Preuzeto s https://www.teradata.com/Glossary/What-is-Semi-Structured-Data (09.06.2022.)

Popis slika

Sažetak

Tehnologija velikih količina podataka i digitalni arhivi

U ovom radu razmatraju se tehnologija velikih količina podataka ili popularnije “big data“ tehnologija i golemi digitalni arhivi koji danas postoje. Pojam „big data” danas je jedan od najvažnijih u području informacijskih znanosti te obuhvaća sve više grana znanosti i tehnologije. „Big data“ je tehnologija koja služi za prikupljanje, obradu i analizu velikih količina podataka. Prikupljeni podaci su strukturirani i nestrukturirani te pristižu velikom brzinom što ih čini složenim za analizu putem normalnih tehnologija. Danas se za rad s „big data" tehnologijom traže stručnjaci sa znanjem programiranja, baza podataka, različitih tipova analize, statistike, vizualizacije podataka, algoritama itd. Cilj ovog završnog rada jest primarno razraditi što je sve to „big data“ tehnologija, definirati pojam i njegove karakteristike, objasniti kako funkcionira obrada i pohrana podataka te osvrt na alate i platforme koje se danas koriste. Također, u radu se opisuju organizacija i princip velikih digitalnih arhiva i pohrane podataka.

Ključne riječi: velika količina podataka, tehnologija, analiza, organizacija, arhivi

Summary

Big data technology and digital archives

This thesis analyzes big data technology and large-scale digital archives. The term „big data“ is of utmost importance in information sciences and encompasses more branches of science and technology. Big Data is a technology used to collect process and analyze large amounts of data. The collected data are structured and unstructured and arrive at high speed, making them difficult to analyze through traditional technologies. Today, experts with knowledge of programming, databases, different types of analysis, statistics, data visualization, algorithms etc., are required to work with big data technology. This thesis aims to investigate what big data technology is, define the term and its characteristics, explain how data processing works and provide a review of the tools and platforms used today. In addition, the thesis describes the organization and principle of extensive digital archives and data storage.

Keywords: big data, technology, analysis, organization, archives