Primljen: 16.03.2023.
Prihvaćen: 11.05.2023.

Stručni rad
UDK: 004.65

Veliki podatci

Big Data

Nenad Breslauer
Međimursko veleučilište u Čakovcu, Bana J. Jelačića 22a, 40000 Čakovec, Hrvatska
e-mail: nenad.breslauer@mev.hr

Sažetak: *U prvom dijelu rada definiran je pojam „veliki podatci“. Zatim je dan pregled upotrebe i izvora velikih podataka. Spomenute su najčešće tehnologije koje se koriste u radu s velikim skupovima podataka. U drugom dijelu rada definiran je pojam strojnoga učenja te su opisane tri vrste strojnog učenja: podržano učenje, nadzirano učenje i nenadzirano učenje. Dat je pregled najpopularnijih tehnologija koje se koriste za strojno učenje i njihove mogućnosti. U zaključku su sumirane osnovne misli rada.*

Veliki podatci (eng. Big data) je popularni termin koji opisuje eksponencijalni rast dostupnosti strukturiranih i nestrukturiranih podataka. Veći broj podatka pridonosi točnijoj analizi. S obzirom na to „veliki podatci“ imaju veliko značenje za samo poslovanje i društvo. [1] Točnijom analizom dobivamo pouzdanije odluke što ujedno donosi veću efikasnost, smanjuje troškove i rizik poslovanja.

Veliki podatci su idealno rješenje za analizu strukturiranih, nestrukturiranih i polustrukturiranih podataka koji dolaze s različitih izvora.

Ključne riječi: *Big data, strojno učenje, Hadoop*

Abstract: *In the first part of the paper, the term "big data" is defined. Then an overview of the use and sources of big data is given. The most common technologies used in working with large data sets are mentioned. In the second part of the paper, the term machine learning is defined and three types of machine learning are described: supported learning, supervised learning and unsupervised learning. An overview of the most popular technologies used for machine learning and their capabilities is given. In the conclusion, the basic ideas of the work are summarized.*

Big data is a popular term that describes the exponential growth in the availability of structured and unstructured data. A larger number of data contributes to a more accurate analysis. Considering this, "big data" has great significance for business and society itself. [1] With a more accurate analysis, we get more reliable decisions, which at the same time brings greater efficiency, reduces costs and business risk.

Big data is an ideal solution for analyzing structured, unstructured and semi-structured data coming from different sources.

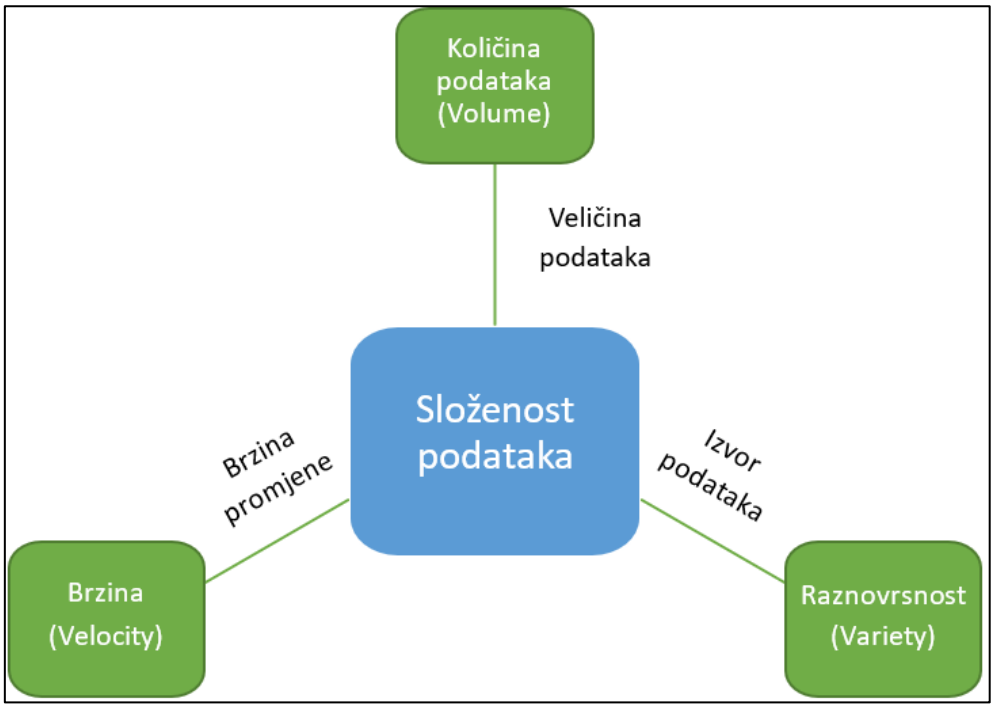
Keywords: *Big data, machine learning, Hadoop*

1. Uvod

Pojam veliki podatci odnosi se na kolekcije podataka koje karakteriziraju tri osobine koje se označavaju kao 3V, što je akronim za engleske riječi: volume, velocity i variety. [2]

Ove tri karakteristike velikih podataka predstavljaju njihove temeljne razlikovne osobine u odnosu na tradicionalne podatke.

Slika 1. *3V Veliki podaci*



Izvor: Autor

Neki autori ističu potrebu da koncept 3V proširi dodatni v-ovima poput: vizije, verifikacije, validacije ili pak varijabilnosti i vjerodostojnosti.

Količina podataka *(eng. Volume)* pojam je koji se mijenja iz godine u godinu. Memorije diskova su sve veće, a cijena je sve manja za istu količinu memorije. Povećanju količine podataka pridonose mnogi faktori kao što su podatci koji su skladišteni godinama, koji se konstantno generiraju na društvenim mrežama itd. U prošlosti su prekomjerne količine podataka stvarale problem oko skladištenja, ali rastom kapaciteta memorije za pohranu i padom cijene to više ne predstavlja problem. Pošto se radi o velikim količinama podataka potrebno ih je analizirati i odrediti važnost podataka iz te gomile.

Brzina *(eng. Velocity)* se odnosi na veliku brzinu nastajanja podataka i brzinu obrade sakupljenih podataka. Podatci pristižu neviđenom brzinom i moraju se pravovremeno obraditi. Brzom obradom podataka tvrtke bi mogle dobiti dodatne informacije koje su im potrebne za donošenje odluka. Važnost leži u povratnoj informaciji, te se razmatra i obrađivanje podataka i tijekom njihova prikupljanja.[4]

Raznovrsnost *(eng. Variety)* govori da podatci u današnje vrijeme dolaze u različitim formatima. Tako imamo strukturirane formate podataka koji se nalaze u bazama podataka i sl., te nestrukturirane i polustrukturirane koji čine većinu podataka današnjice. Stoga je bitno se usredotočiti se na sve vrste podataka te ih kombinirati kako bi se povećala njihova vrijednost.

2. Upotreba i izvori velikih podataka

Velike količine podataka mogu dati odgovore na bitna poslovna pitanja. Dobivanje i upotreba velikih količina podataka obećavajuća je vizija koja će organizacijama omogućiti da prikupe i analiziraju relevantne podatke. Dobrom analizom podataka mogu pronaći odgovore kako smanjiti troškove i vrijeme poslovanja, razviti nove proizvode te pametnije prilagoditi

ponudu i samo odlučivanje unutar poslovanja. Veliki podatci mogu dati (točnije) odgovore na bitna poslovna pitanja.

Prema istraživanju Microsofta, provedenom na istraživanju 6200 malih i srednjih poduzeća u 20 zemlja Europe, ustanovljeno je da će poduzeća koje koriste analizu velikih podataka kod odlučivanja, lansirati nove usluge i proizvode te ih širiti na ostala tržišta. [5]

Te će velike podatke usmjeriti k tri ključne značajke:

- koristit će podatke da bi dobili nove klijente
- koristit će podatke kako bi pripremili svoj tim za nove izazove poslovanja (osposobiti tim za korištenje novih tehnologija i podataka kako bi se doprinijelo rastu poslovanja.)
- koristit će podatke kako bi ostvarili svoju priliku prije konkurencije (70 % poduzeća smatra da je upotreba velikih podataka povećala njihovu sposobnost inoviranja) [5]

Podatci su za informacijsko društvo veoma bitni jer bez njih nema inovacija o kojima se danas ovisi. Veliki podatci su središte znanosti i poslovanja. Vrijednost velikih podataka leži u mogućnostima njihove primjene. Moguće ih je primijeniti, na više načina: novom uporabom „starih" podataka, spajanjem različitih skupova podataka, višenamjenskom uporabom podataka.

Izvori podataka mogu se preuzeti s web stranice poput Infochimps.org, theinfo.org ili Amazon Web Services gdje se podatci mogu preuzeti besplatno ili po određenoj cijeni. Prije su se pokušavali izvesti složeni algoritmi koji su trebali biti izuzetno dobri kako bi se došlo do rezultata, a danas se u tu svrhu koriste podaci koji mogu biti znatno bolji. Javno dostupni izvori podataka, kao što su The World Factbook – CIA [6] ili European Union Open Data Portal [7] predstavljaju izvor ogromne količine podataka.

3. Tehnologije za rad s velikim skupovima podataka

Tehnologije, osim za prikupljanja velikih količina podataka, omogućavaju i izvlačenje vrijednosti samih podataka kao i njihovo razumijevanje. Potrebno je pronaći tehnologije koje će imati mogućnosti da obrade veliku količinu podataka bez prevelikih troškova. Prvi koji su napravili proboj u tehnologiji su Yahoo!, Google i Facebook.

Apache Hadoop je projekt visoke razine koji je napisan u Javi i koji je dizajniran upravo u svrhu izvođenja operacija nad velikim podacima. Hadoop je implementiran u mnoge velike korporacije kao što su: Apple, Facebook, HP, Netflix i drugi. Hadoop je omogućio otkrivanje informacija pomoću skeniranja velikih podataka kroz visoku skalabilnost i distribuirani batch sustav za obradu. Hadoop se promatra kroz dva dijela : HDFS (Hadoop Distributed File System) i MapReduce. Uz ta dva dijela bitno je spomenuti i ove projekte koji su povezani s Hadoopom: Apache Avro (služi za sterilizaciju podataka), Hbase i Cassandra (baze podataka), Hive (pruža ad hoc upite koji su slični SQL upitima), Chukwa, Mahout ,Pig , ZooKeeper i mnogi drugi. [8]

3.1. Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) pohranjuje podatke u tzv. blokove i kopira te blokove na druge servere u Hadoop klasteru, podatke dijeli na manje blokove koji se dijele na servere unutar klastera. Zadana veličina bloka u Hadoopu je 64MB, a za veće podatke moguće je koristit i blokove od 128 MB. Svaki blok se pohranjuje na tri servera kako bi se povećala pouzdanost i sigurnost. Cijela logika smještanja podataka se odvija zahvaljujući NameNode - u. NameNode se brine umjesto čovjeka, gdje će se smjestiti podatci. Hadoop kontaktira NameNode, koji pronalazi gdje su podatci koji se traže i šalje ih dalje aplikaciji koja ih pokreće lokalno na tim nodovima. Stvarni podatci koji se analiziraju MapReduceom ne prolaze kroz NameNode, već se on koristi samo za upravljanje meta podatcima koji opisuju gdje se podatci nalaze.

3.2. MapReduce

Riječ je o programskoj paradigmi koja omogućuje ogromnu skalabilnost kroz tisuće servera unutar Hadoop klastera. MapReduce se odnosi na posao koji se dijeli na manje dijelove ili zadatke. Aplikacija dodjeljuje posao Hadoop klasteru koji pokreće JobTracker. Kako bi saznao gdje su sve podatci koji trebaju nalaze unutar klastera te se posao dijeli na manje dijelove tj. zadatke. Također, bitan je i TaskTracker kojem je posao pratiti status svakoga taska ili zadatka. Zadatak će se dodijeliti novom Nodu unutar klastera ako JobTrackeru stavi status da nije uspio. [9,10]

3.3. Pig

Pig je skriptni jezik koji se upotrebljava se za pisanje kompleksnih "MapReduce" transformacija korištenjem jednostavne sintakse bez potrebe za znanjem Jave. Pig Latin (Pigov jednostavni jezik za skriptiranje, sličan SQL-u) definira niz transformacija na skupu podataka kao što su agregacije, spajanja i sortiranja podataka. Pig prevodi skripte pisane Pig Latin jezikom u "MapReduce" kako bi ih bilo moguće izvršavati na Hadoopu. Pig Latin se ponekad proširuje upotrebom funkcija definiranih sa strane korisnika koje može napisati u Javi ili nekom skriptnom jeziku te zatim pozvati direktno pomoću Pig Latin jezika. [11]

3.4. Hive

Kako bi se olakšao cijeli proces još više, stvoren je Hive, koji ima slično okruženje SQL-u. HQL – Hive Query language koji ima određena ograničenja, ali je i dalje veoma koristan. Njegove naredbe se dijele na MapReduce poslove i izvršavaju se kroz Hadoop klaster. Hive se bazira na Hadoop i MapReduce operacijama, ali postoje neke razlike. Zbog toga što je Hadoop napravljen za sekvencijalno skeniranje, očekuju se upiti kojima treba dugo da se izvrše. Ukoliko nam treba veoma brz izvještaj onda ovo predstavlja problem. [12]

3.5. ZooKeeper

Radi se usluzi koja osigurava sinkronizaciju kroz klaster. Ukoliko imamo i manju količinu servera nužna je centralizacija kad govorimo o upravljanju, a pogotovo kad se radi o velikom broju servera. ZooKeeper server čuva kopije stanja cijelog sustava i svaki klijent komunicira jednog ZooKeeper servera (može ih biti više), kako bi vratio ili nadogradio informaciju o sinkronizaciji. [13]

4. Strojno učenje

Strojno učenje (*eng. Machine Learning)* je postupak po kojem se tradicionalna analiza razlikuje od obične analize podataka. Definicija strojnoga učenja predstavlja dizajn računalnih algoritama koji koriste iskustvo iz prošlosti prilikom donošenja budućih odluka; to je studija programa koji uče iz podataka. Cilj strojnoga učenja je generalizacija, odnosno sposobnost algoritma da njegova primjena bude što točnija na novim podatcima/zadatcima nakon izvjesnoga treniranja na prethodnim podatcima. [14]

Postoje tri vrste strojnoga učenja [15]: 1. nadzirano učenje (*eng. Supervised learning*), 2. nenadzirano učenje (*eng. Unsupervised learning*) i 3. podržano (ojačano) učenje (eng. Reinforcement learning). Nadzirnom učenje podatci za učenje su u obliku (ulaz, izlaz). Cilj učenja jest pronaći preslikavanje $\hat{y} = f(x)$ s ulaza na izlaz. „Ukoliko je diskretna vrijednost, tada se problem naziva klasifikacija, a ukoliko je kontinuirana vrijednost, tada se naziva regresija. Kod nenadziranoga učenja, u skupu za učenje, nalaze se podatci bez ciljne vrijednosti. Cilj nenadziranoga učenja jest pronaći pravilnosti u podatcima. Podržano ili ojačano učenje jest učenje optimalne strategije na temelju pokušaja s odgođenom nagradom.“ Tipične primjene podržanoga učenja su igranje igara, robotika, upravljanje i više agentski sustavi. [16]

4.1. Alati za strojno učenje

Postoji mnogo alata za izvođenje postupka strojnoga učenja, a pogotovo su zanimljivi alati čije je korištenje besplatno.

Weka je platforma za strojno učenje koja sadrži najpoznatije algoritme strojnoga učenja. Uključuje cijeli niz alata koji olakšavaju mnoge aktivnosti strojnoga učenja kao što je priprema podataka i vizualizacije rezultata. Rudarenje podatcima je proces pronalaženja uzoraka (*engl.patterns*) među podatcima. Weka uključuje pripremu ulaznih podataka, statističku evaluaciju procesa učenja i vizualizaciju ulaznih podatka i rezultata strojnoga učenja. [17]

RapidMiner je vodeći sustav otvorenoga koda za dubinsku analizu podataka i otkrivanje znanja. RapidMiner je Java okružje za strojno učenje, rudarenju tekstualnih podataka i otkrivanju znanja u bazama podataka. Koristi se za dubinsku analizu podataka, odlikuje se jednostavnim grafičkim sučeljem. [18]

Matlab je programski jezik visoke razine namijenjen za tehničke proračune. Objedinjuje računanje, vizualizaciju i programiranje u lako uporabljivoj okolini u kojoj su problem i rješenje definirani poznatom matematičkom notacijom.

Matlab je koncipiran kao proširiv programski paket, osim osnovnoga paketa moguće je nabaviti i dodatne module specijalizirane za rad u područjima kao što su automatsko upravljanje (Control System Toolbox), obrada signala (Signal Processing Toolbox) ili simulacija neuronskih mreža (Neural Network Toolbox).

Osnovna karakteristika matlab-a je da nalazi numerička rješenja različitih vrsta matematičkih problema, od onih najjednostavnijih do izrazito složenih. Druga bitna karakteristika matlab-a je širok spektar mogućnosti za grafičko prikazivanje. matlab generira grafike specifične po boji, osvijetljenosti, vrsti linija i tekstualnom sadržaju. [19]

5. Zaključak

U ovom radu su se prezentirali pojmovi veliki podatci (*engl. Big Data*) i strojno učenje (*engl. Machine Learning*) i njihove najčešće tehnologije.

Trenutačno u svijetu ima više podataka, odnosno informacija nego ikada prije, a broj informacija svakim danom se povećava. Veliki podatci mogu dovesti do određenih zaključaka zbog velike količine podataka, a do kojih ne bi bilo moguće doći da je ta skala podataka mala. Uz pomoć tih podataka moguće je doći do novih vrijednosti koja mijenjaju tržišta, organizacije, odnosno način na koji čovjek živi i djeluje. Transformacija gotovo svih segmenata ljudske stvarnosti u podatke novost je za većinu ljudi u sadašnjosti. Svijest o rastu Velikih podataka i pretpostavka da postoji mjerljiva komponenta u gotovo svemu što činimo, dovode do toga te da su dobiveni podatci ogromni izvor znanja, značajno će utjecati na našu stvarnosti.

Strojno učenje je zasigurno jedna od znanstvenih disciplina koja će u budućnosti najviše napredovati. Strojno učenje je idealno za iskorištavanje mogućnosti skrivenih u velikim podatcima. Što više podataka posjedujemo za strojno učenje, to se može bolje učiti, odnosno njegova primjena na novim podatcima bit će kvalitetnija.

Literatura

[1] Berman, J. J. (2013), Principles of Bigdata: preparing, sharing, and analysing complex information, Elsevier, Morgan Kaufman, Amsterdam.
[2] Chun-Hsin Wu, Pu Hsu; (2015), Cost-Effective and Reliable Cloud Storage for Bigdata, Dept. Computer Science and Information Engineering National University of Kaohsiung Kaohsiung, Taiwan,
[3] Hilbert, M., Lopez, P. (2012), „How to Measure the World's Technological capacity to Communicate, Store, and Compute Information Part I: Results and Scope “u International Journal of Communication 6, 956-979, dostupno na:

http://ijoc.org/index.php/ijoc/article/view/1562/742 (10.1.2023.)
[4] Analitika BD. http://www.webopedia.com/TERM/B/big_data_analytics.html (12.1.2023.)
[5] https://news.microsoft.com/europe/2016/04/20/go-bigger-with-big-data/ (21.12.2022.)
[6] https://www.cia.gov/the-world-factbook/ (21.12.2022.)
[7] https://data.europa.eu/en (1.3.2023.)
[8] https://hadoop.apache.org/ (4.1.2023.)
[9] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html (4.1.2023.)
[10] Sheshasaayee, A., Lakshmi J. V. N.: A Theoretical Model for Big Data Analytics using Machine Learning Algorithms
[11] http://pig.apache.org/ (4.1.2023.)
[12] http://hive.apache.org/ (4.1.2023.)
[13] http://zookeeper.apache.org/ (4.1.2023.)
[14] Xiang, J., Westerlund M., Sovilj D., Pulkkis G. (2014.): Using Extreme Learning Machine for Intrusion Detection in a Big Data Environment.
[15] http://www.zemris.fer.hr/~yeti/studenti/Seminar_2/2012/Knezevic_Karlo/Seminar%5B2012%5DKne_evi__Karlo.pdf (10.3.2022.)
[16] https://strojnoucenje.takelab.fer.hr/ (10.2.2023.)
[17] https://hr.myservername.com/weka-tutorial-how-download (2.2.2023.)
[18] https://rapidminer.com/ (14.1.2023.)
[19] http://www.mathworks.com (14.1.2023.)