Nestrukturirani podatci i novi modeli obrade, pretraživanja i reprezentacije

Galić, Ljerka

Sveučilište Josipa Jurja Strossmayera u Osijeku
Ekonomski fakultet u Osijeku
Preddiplomski studij

Ljerka Galić

NESTRUKTURIRANI PODATCI I NOVI MODELI OBRADE, PRETRAŽIVANJA I REPREZENTACIJE

Završni rad

Osijek, 2020.

Sveučilište Josipa Jurja Strossmayera u Osijeku
Ekonomski fakultet u Osijeku
Preddiplomski studij

NESTRUKTURIRANI PODATCI I NOVI MODELI OBRADE, PRETRAŽIVANJA I REPREZENTACIJE

Završni rad

Kolegij: Upravljanje informacijskim resursima
JMBAG: 0010221787
email:ljerkagalic924@gmail.com

Mentor: prof.dr.sc. Josip Mesarić

Osijek, 2020.

Josip Juraj Strossmayer University of Osijek

Faculty of Economics in Osijek

Undergraduate Study, Business informatics

UNSTRUCTURED DATA AND NEW MODELS OF PROCESSING, SEARCH AND REPRESENTATION

Final paper

Osijek, 2020.

IZJAVA

O AKADEMSKOJ ČESTITOSTI, PRAVU PRIJENOSA INTELEKTUALNOG VLASNIŠTVA, SUGLASNOSTI ZA OBJAVU U INSTITUCIJSKIM REPOZITORIJIMA I ISTOVJETNOSTI DIGITALNE I TISKANE VERZIJE RADA

1. Kojom izjavljujem i svojim potpisom potvrđujem da je završni (navesti vrstu rada: završni / diplomski / specijalistički / doktorski) rad isključivo rezultat osobnoga rada koji se temelji na mojim istraživanjima i oslanja se na objavljenu literaturu. Potvrđujem poštivanje nepovredivosti autorstva te točno citiranje radova drugih autora i referiranje na njih.
2. Kojom izjavljujem da je Ekonomski fakultet u Osijeku, bez naknade u vremenski i teritorijalno neograničenom opsegu, nositelj svih prava intelektualnoga vlasništva u odnosu na navedeni rad pod licencom *Creative Commons Imenovanje – Nekomercijalno – Dijeli pod istim uvjetima 3.0 Hrvatska.* (cc) BY-NC-SA
3. Kojom izjavljujem da sam suglasan/suglasna da se trajno pohrani i objavi moj rad u institucijskom digitalnom repozitoriju Ekonomskoga fakulteta u Osijeku, repozitoriju Sveučilišta Josipa Jurja Strossmayera u Osijeku te javno dostupnom repozitoriju Nacionalne i sveučilišne knjižnice u Zagrebu (u skladu s odredbama Zakona o znanstvenoj djelatnosti i visokom obrazovanju, NN br. 123/03, 198/03, 105/04, 174/04, 02/07, 46/07, 45/09, 63/11, 94/13, 139/13, 101/14, 60/15).
4. izjavljujem da sam autor/autorica predanog rada i da je sadržaj predane elektroničke datoteke u potpunosti istovjetan sa dovršenom tiskanom verzijom rada predanom u svrhu obrane istog.

Ime i prezime studenta/studentice: Ljerka Galić

JMBAG: 0010221787

OIB: 21163638297

e-mail za kontakt: ljerkagalic924@gmail.com

Naziv studija: Preddiplomski sveučilišni studij Poslovna Informatika

Naslov rada: Nestrukturirani podatci i novi modeli obrade, pretraživanja i reprezentacije

Mentor/mentorica završnog rada: prof. dr. sc. Josip Mesarić

U Osijeku, 7. 09 2020. 2020. godine

Potpis Ljerka Galić

Sažetak

U današnje vrijeme na raspolaganju je velika količina podataka a većina se nalazi u nestrukturiranom ili polustrukturiranom obliku. Slabo strukturirani i nestrukturirani podatci su podatci koji se ne mogu obrađivati klasičnim sustavima za obradu podataka. Ideja je na neki ih način pretvoriti u strukture koje se mogu analizirati i pretraživati po smislenim algoritmima. Nestrukturirani i nepohranjeni podatci smanjuju iskoristivost, mogućnost analize i korištenje tih dokumenata stoga je potrebno pronaći način za obradu i korištenje takvih podataka. Upravljanje nestrukturiranim podatcima i sadržajem važan je zahtjev u poslovanju bilo kojeg poslovnog subjekta. Strukturirana je samo manja količina informacija koje se koriste u transakcijskim poslovnim sustavima kao što su ERP ili CRM. Pod nestrukturiranim dokumentima podrazumijevaju se svi dokumenti za koje bilo kakav podatak o dokumentu možemo naći samo otvaranjem (čitanjem) dokumenta. Veći dio dokumenata u bilo kojoj bazi podataka je nestrukturiran i pohranjen u različitim formatima (tekstovi, razne elektroničke forme: od e-mail poruka, preko MS Word dokumenta, sve do web stranica i multimedijalnih sadržaja), što predstavlja problem u korištenju i razumijevanju takvih podataka, posebno ukoliko se ti podatci koriste za poslovanje tvrtke. Stoga je potrebno pronaći rješenje za nestrukturirane podatke. Postoje različiti alati i baze podataka koje pomažu u rješavanju tog problema. U radu su ukratko opisani alati za obradu podataka Mozenda, Octoparse i Import.io, također su opisane NoSQL baze podataka te Hadoop softver koji je pojašnjen i na primjeru. Time su navedena samo neka rješenja problema velike količine nestrukturiranih podataka. Rad predstavlja sintetizirani pregled aktualnih modela i metoda obrade masovnih te slabostrukturiranih podatka

Ključne riječi: podatci, informacije, nestrukturirani podatci, informacijski sustavi, baze podataka, analiza podataka

Summary

Nowadays, a large amount of data is available and most are in unstructured or semi-structured form. Poorly structured and unstructured data are data that cannot be processed by classical data processing systems. The idea is to somehow turn them into structures that can be analyzed and searched by meaningful algorithms. Unstructured and unsaved data reduce the usability, analysis and use of these documents, so it is necessary to find a way to process and use such data. Managing unstructured data and content is an important requirement in the business of any business entity. Only a small amount of information used in transactional business systems such as ERP or CRM is structured. Unstructured documents are all documents for which any information about the document can be found only by opening (reading) the document. Most of the documents in any database are unstructured and stored in various formats (texts, various electronic forms: from e-mails, through MS Word documents, all the way to web pages and multimedia content), which is a problem in using and understanding such data, especially if that data is used for the company's business. Therefore, it is necessary to find a solution for unstructured data. There are various tools and databases that help solve this problem. The paper briefly describes the data processing tools Mozenda, Octoparse and Import.io, as well as NoSQL databases and Hadoop software, which is explained by example. This lists only some solutions to the problem of large amounts of unstructured data. The paper presents a synthesized overview of current models and methods of processing mass and poorly structured data.

Keywords: data, information, unstructured data, information system, database, data analysis

Sadržaj

Popis tablica:

Popis slika:

1.Uvod

Problemi s obradom masovnih podataka različitih formata i strukture

Unaprijeđivanje procesa upravljanja informacijama postaje neizbježno u tvrtkama i organizacijama svih veličina jer one predstavljaju njihovu značajnu intelektualnu imovinu. Mnoge se organizacije (financijske ustanove, osiguravateljske kuće, vladine tvrtke i agencije, zdravstvene ustanove) u pružanju svojih usluga još uvijek oslanjaju na poslovne procese temeljene na ručnoj ili poluatomatiziranoj obradi i rukovanju papirnatim dokumentima te višekratnim osobnim evidencijama i kopijama tih dokumenata.(Senso-is.hr)[1]. Upravljanje nestrukturiranim podatcima i sadržajem važan je zahtjev u poslovanju bilo kojeg poslovnog subjekta. Vrlo je važno razumjeti izvor podataka koji je koristan za poduzeće. Može se koristiti jedan ili više izvora podataka za prikupljanje informacija relevantnih za poslovanje. Prikupljanje podataka iz slučajnih izvora nikada nije dobra ideja, jer se time podatci mogu oštetiti ili čak izgubiti. Strukturirana je samo manja količina informacija koje se koriste u transakcijskim poslovnim sustavima kao što su ERP, CRM. Pod nestrukturiranim dokumentima podrazumijevaju se svi dokumenti za koje bilo kakav podatak o dokumentu možemo naći samo otvaranjem odnosno čitanjem dokumenta. Veći dio dokumenata u bilo kojoj bazi podataka je nestrukturiran i pohranjen u različitim formatima (papir, razne elektroničke forme: od e-mail poruka, preko MS Word dokumenta, sve do web stranica i multimedijalnih sadržaja) što predstavlja problem u korištenju i razumijevanju takvih podataka posebno ukoliko se ti podatci koriste za poslovanje tvrtke. Za bolje razumijevanje nestrukturiranih podataka potrebno je obrazložiti razliku između strukturiranih i nestrukturiranih podataka i njihove karakteristike. Strukturirani podaci obično se nalaze u relacijskim bazama podataka (RDBMS). U polja se pohranjuju telefonski brojevi, brojevi socijalnog osiguranja ili poštanski brojevi (Taylor, 2018)[2]. Čak su i tekstualni nizovi promjenjive duljine poput imena sadržani u zapisima, što ih čini jednostavnim za pretraživanje. Podaci se mogu generirati odnosno pretraživati sve dok su podaci stvoreni unutar RDBMS strukture. U ovom se formatu može pretraživati i pomoću upita generiranih od strane čovjeka te putem algoritama pomoću vrste podataka i naziva polja, poput abecednog ili numeričkog,

[1] Upravljanje dokumentima, http://www.senso-is.hr/upravljanje-dokumentima.aspx , Pristupljeno 5.svibnja.2020

[2] Taylor, C., (2018), Structured vs. Unstructured Data, https://www.datamation.com/big-data/structured-vs-unstructured-data.html, Pristupljeno 07.07.2020.

valute ili datuma. Uobičajene aplikacije za relacijske baze podataka sa strukturiranim podacima uključuju zrakoplovne rezervacijske sustave, kontrolu zaliha, transakcije prodaje i ATM aktivnosti. Strukturirani jezik upita (SQL) omogućuje upite na ovu vrstu strukturiranih podataka unutar relacijskih baza podataka. Neke relacijske baze podataka pohranjuju ili upućuju na nestrukturirane podatke, kao što su aplikacije za upravljanje odnosima s klijentima (CRM). Nestrukturirani podaci su u osnovi sve ostalo. Pojam nestrukturirane informacije odnosi se na informacije koje ili nemaju unaprijed definiran model podataka ili nisu organizirane na unaprijed definiran način (Analitycs, 2018.). Nestrukturirani podaci imaju unutarnju strukturu, ali nisu strukturirani putem unaprijed definiranih modela podataka ili sheme. Može biti tekstualna ili netekstualna, te generirana od ljudi ili stroja. Također se može pohraniti u nerelacijsku bazu podataka poput NoSQL. Pored očite razlike između pohrane u relacijskoj bazi podataka i pohrane izvan jedne, najveća razlika je jednostavnost analize strukturiranih podataka u odnosu na nestrukturirane podatke. Zreli alati za analitiku postoje za strukturirane podatke, ali analitički alati za rudarstvo nestrukturiranih podataka se tek razvijaju. Korisnicimogu pokrenuti jednostavno pretraživanje sadržaja preko tekstualnih nestrukturiranih podataka. No, nedostatak uredne unutarnje strukture smanjuje svrhu tradicionalnih alata za obradu podataka, a poduzeća time gube veliku količinu potencijalno korisnih informacija. Nestrukturirani podaci čine 80% i više podataka o poduzeću i rastu po stopi od 55% i 65% godišnje. Bez korištenja alata za analizu ovih ogromnih podataka, organizacije ostavljaju ogromne količine vrijednih podataka neiskorištenima.

2. Svrha i ciljevi rada

Istražiti aktualne modele, metode i njihove ključne karakteristike za prikaz, obradu i pretraživanje nestrukturiranih podataka. Pojasniti razliku između strukturiranih i nestrukturiranih podataka te pobliže opisati na primjerima. Opisati pojedine alate i modele za obradu i pretraživanje podataka, pojasniti njihove prednosti i nedostatke te u kojim slučajevima ih je najbolje koristiti. Na kraju rada pojasniti će se primjena ovakvih modela i općenito nestrukturiranih podataka na stvarnim primjerima. U ovom radu obrađivat će se osnovne karakteristike alata Import.io, Mozenda i Octoparse, te pobliže opisati rad NoSQL baze podataka i Hadoop sustava. Glavni cilj je objasniti važnost nestrukturiranih podataka i opisati na koje se sve načine mogu obrađivati i pohranjivati velike baze podataka.

3. Metodologija istraživanja

Istraživanje se temelji na analizi i sintezi aktualne literature i studiju slučaja raspoloživih metodologija za upravljanje nestrukturiranim podatcima. U pronalasku literature o nestrukturiranim podatcima te modelima i alatima za obradu podataka korišten je internet i internetski članci te web stranice konkretnih modela i alata za obradu podataka. Također je bilo potrebno proučavanje opisanih alata i modela za obradu i pohranu podataka kako bi isti mogli biti opisani u radu.

4. Prevođenje nestrukturiranih podataka u strukturirane

Glavna razlika između strukturiranih i nestrukturiranih podataka je u tome što se strukturirani podatci lakše pretražuju, dok su nestrukturirani podatci svi ostali podatci koji se ne pretražuju na jednostavan način jer se nalaze u različitim formatima. Stoga je takve podatke teže analizirati i iskoristiti za daljnje donošenje odluka. Postoje razlike između jednostavnosti analize strukturiranih podataka i zahtjevnije analize nestrukturiranih podataka. Strukturirana analiza podataka zreli je proces i tehnologija. Nestrukturirana analitika podataka početna je industrija s puno novih ulaganja u istraživanje i razvoj, ali nije zrela tehnologija. Pitanje strukturiranih podataka i nestrukturiranih podataka unutar poslovnog subjekta odlučuje trebaju li ulagati u analitiku nestrukturiranih podataka i ako je moguće objediniti ih u bolju poslovnu inteligenciju. U tablici 1 prikazane su osnovne razlike i karakteristike strukturiranih i nestrukturiranih podataka.

Tablica 1 Razlike između strukturiranih i nestrukturirsnih podataka Izvor: Structured vs. Unstructured Data , By Christine Taylor, Posted March 28, 2018. https://www.datamation.com/big-data/structured-vs-unstructured-data.html, prevedeno i prilagođeno od strane autorice

	Strukturirani podatci	Nestrukturirani podatci
Karakteristike	• Prethodno definirani modeli podataka • Tekstualni podatci • Jednostavno pretraživanje	• Modeli podataka koji nisu prethodno definirani • Mogu biti tekstualni,slike,audio zapisi i videozapisi i drugi • Teško ih je pretraživati
Nalazi se u	• Relacijske baze podataka • Skladišta podataka	• Aplikacije • NoSQL baze podataka • Skladišta podataka • Data lakes(jezera podataka)

Upravljanje	Čovjek ili stroj	Čovjek ili stroj
Uobičajna primjena	• Sustavi za rezervacije zarakoplovnih kompanija • Kontrola inventara • CRM sustavi • ERP sustavi	• Word procesor • Prezentacijski softveri • Čitači e-pošte • Alati za pregled i obradu
Primjeri	• Datumi • Telefonski brojevi • Brojevi kreditnih kartica • Imena klijenata • Adrese • Transakcijske informacije	• Tekstualni dokumenti • Izvješća • Email poruke • Audio i videozapisi • Slike

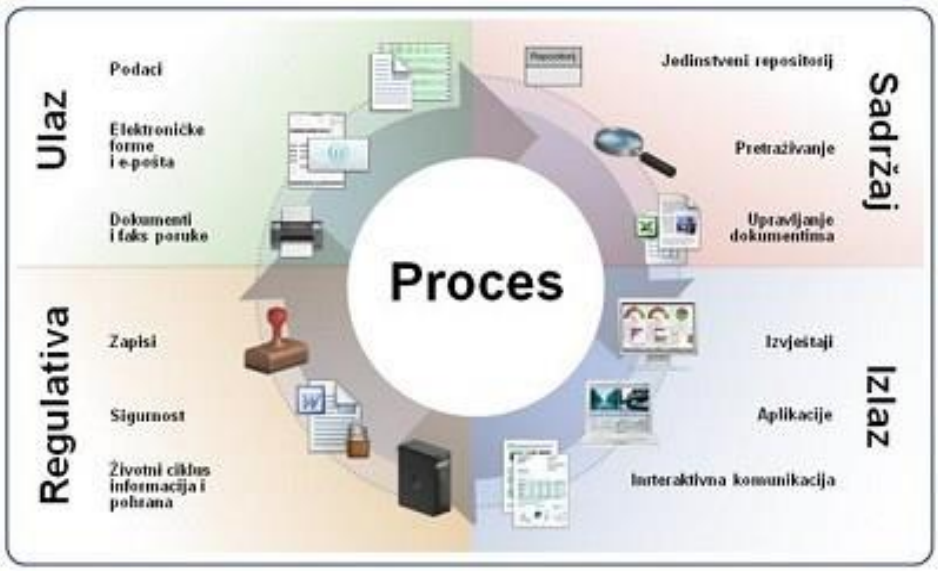
Pored čisto strukturiranih i nestrukturiranih podataka postoje i polustrukturirani podatci. To su podaci koji „održavaju interne oznake i oznake koje identificiraju odvojene podatkovne elemente, što omogućava grupiranje podataka i hijerarhiju. Dokumenti i baze podataka također mogu biti polustrukturirani. E-pošta vrlo je čest primjer polustrukturirane vrste podataka. izvorni metapodaci e-pošte omogućuju klasificiranje i pretraživanje ključnih riječi bez dodatnih alata. Osnovna meta baza podataka e-pošte ne dozvoljava pretraživanje ključnih riječi bez korištenja dodatnih naprednijih alata“ (Taylor, 2018.)[3]. Prevođenje nestrukturiranih podataka u strukturirane je složen proces, posebno ako se radi o velikoj bazi podataka(Big data). Neka od rješenja su alati za prevođenje nestrukturiranh podataka koji mogu raditi na principu web scrapinga odnosno ekstrakcijom sadržaja primjer takvih alata su Mozenda i Octoparse. Također alati mogu funkcionirati i na principu obrade i pripreme podataka s interneta. Ovi alati su opisani pobliže u daljnjem nastavku rada. Kao rješenje nagomilanih nestrukturiranih podataka također predstavljaju i NoSql baze podataka koje označavaju „ne samo SQL“ baze podataka i namjenjene su za velike količine podataka. Hadoop softver za pohranu i obradu podataka također može obraditi velike kolčine podataka. Međutim u budućnosti nestrukturiranih podataka teži se ka jedinstvenom rješenju pretvorbe podataka u strukturirane oblike.

Sustavi za upravljanje sadržajima

Kako bi poslovni subjekti iskoristili sve podatke koje imaju na raspolaganju na pravi način potrebno je prevesti nestrukturirane podatke u oblik u kojem ih je najlakše pretraživati i koristiti. Rješenja za upravljanje poslovnim sadržajima, Enterprise Content Management (ECM) omogućuju organizacijama spajanje sadržaja i poslovnih procesa kroz jedinstvenu platformu. ECM je kategorija softvera koja omogućava upravljanje svim nestrukturiranim

[3] Taylor, C, Ibidem

informacijama, odnosno sadržajem, na sustavan i integriran način (Senso-is.hr). Omogućuje organizacijama pohranu dokumenata i informacija te nudi brojne prednosti korištenja ovog sustava. Neke od prednosti su smanjenje pogrešaka, smanjenje arhive dokumenata i njihovog gubitka, brz pristup informacijama, pojednostavljenje poslovnih procesa, bolju kontrolu dokumenata. Osnovni princip rada ovakvog sustava sastoji se od nekoliko ključnih aktivnosti. ECM sustav sve važne dokumente pohranjuje u digitalnom obliku i objedinjuje u jedan repozitorij. U digitalnom obliku svi dokumenti su spremni za čuvanje a kasnije i obradu i pretraživanje, teži se automatiziranom pretraživanju dokumenata. Važno je da sustav ovakve vrste ima visoku razinu zaštite i sigurnosti kako poslovni subjekti ne bi izgubili dokumente. Na slici 1 prikazano je upravljanje dokumentima po različitim fazama.



Slika 1 Upravljanje dokumentima Izvor: http://www.senso-is.hr/upravljanje-dokumentima.aspx , pristupljeno 5.05.2020

5. Alati za prikupljanje, obradu, pretraživanje i prikaz masovnih slabo strukturiranih podataka

Organizacije i poslovni subjekti, u današnjem svijetu temeljenom na podacima, se ne mogu oslanjati samo na podatke iz svojih unutarnjih sustava za donošenje poslovnih odluka i vođenje poslovnih procesa. Potrebno je također koristiti vanjske izvore podataka odnosno alternativne skupove podataka koje im olakšavaju proces pronalaska odgovarajućih informacija i iskorištavanja istih. Ljudski se jezik prilično razlikuje od jezika koji koriste strojevi koji preferiraju strukturirane informacije. Stoga je cilj nestrukturiranog alata za analizu podataka izgraditi most između ova dva vrlo različita jezika. U nastavku će biti opisani neki od alata za obradu i pohranu nestrukturiranih podataka.

5.1. Import.io

Internet je najveće svjetsko spremište podataka i izvor je ogromnog potencijala za nova znanja i vještine. Međutim, takve ogromne baze podataka teško je pretraživati. Import.io omogućuje pretvaranje nestrukturiranih podataka u strukturirani oblik obradom i pripremom web podataka. Posjeduje vlastiti modul za vizualizaciju kako bi poslovni analitičari stekli uvid koji im je potreban, a nudi i aplikacijsko programsko sučelje koje posjeduje potpuni pristup svemu što se može učiniti na njihovoj platformi, omogućuje izravnu integraciju internetskih podataka u vlastite aplikacije. Import.io je alat koji nudi brzu i jednostavnu implementaciju projekata, vizualizaciju podataka putem grafikona i izvješća o podatcima koje su važna za poslovni subjekt. Na slici 2 je prikazan primjer korištenja alata Import.io.

import.io
CHANGE REPORT

amazon.com Mr. Coffee Change Report

ITEMS CHANGED (23)
Previous run Thu Dec 07 2017 16:52:36 GMT+0000 (UTC)
Current run Fri Dec 08 2017 16:52:25 GMT+0000 (UTC)

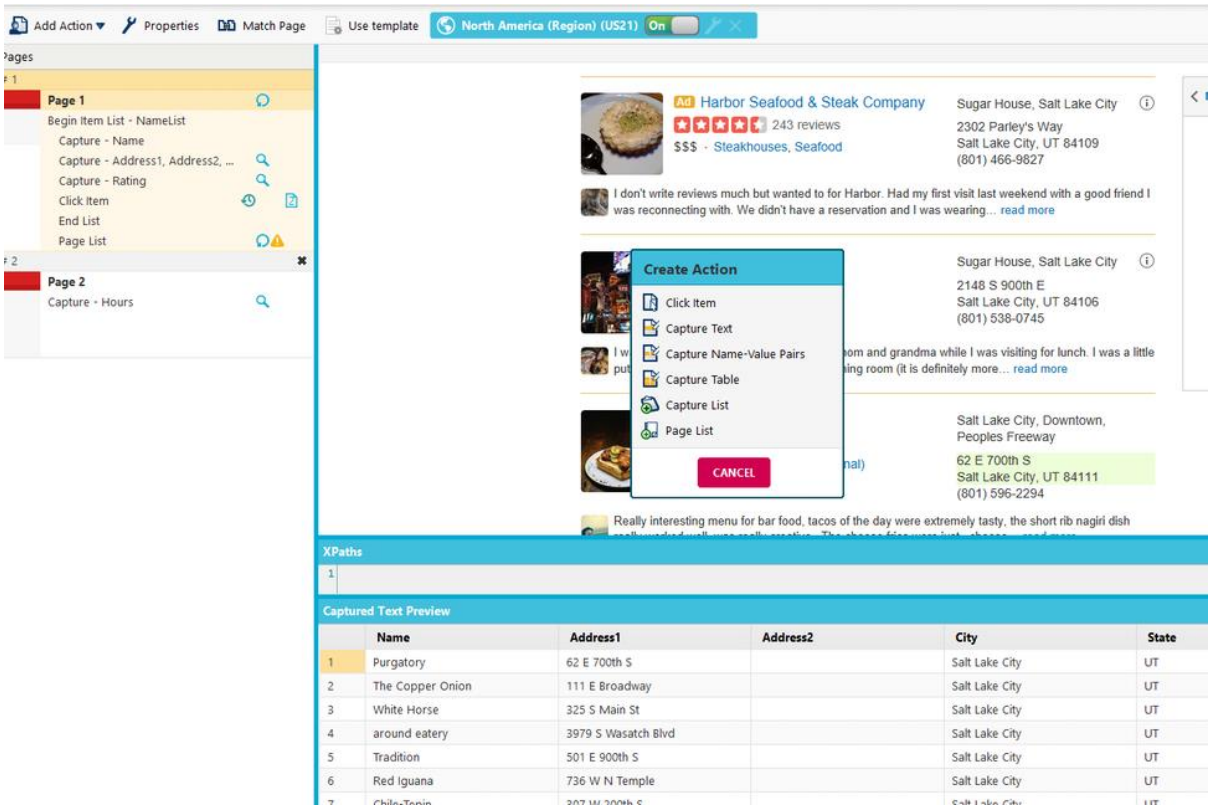
Product Name	Price	
	Previous	Current
Mr. Coffee Ice Tea Glass Pitcher 2.5 QT, BVST-TP23	$ 29 99	$ 29 69
Everyday 12-Replacement Charcoal Water Filters for Mr. Coffee Machines	$ 8 30	$ 8 50
Mr. Coffee Espresso Carafe Assembly w/ Lid, Black 4 Cup	$ 12 58	$ 12 57
Mr. Coffee BVMC-TP1 2-Quart Replacement Pitcher for TM1, TM1P	$ 13 89	$ 15 78
1 X 4-Cup Basket Style Permanent Coffee Filter fits Mr. Coffee 4 Cup Coffeemakers (With Handle)	$ 5 79	$ 6 69
Mr. Coffee 5-Cup Coffee Maker, Black	$ 15 77	$ 18 52
Mr. Coffee Water Filter Replacement Disk, 2 Pack	$ 6 53	$ 7 65
Mr. Coffee 12-Cup Programmable Coffee Maker	$ 49 99	$ 36 84
12-Cup FT Series Replacement Decanter with Drip-Resistant Lid	$ 19 00	$ 18 40
Mr. Coffee 3 Piece Piezas 1.2 Quart Coffee Press	$ 16 57	$ 20 49
Mr. Coffee Basket Coffee Filters, 8-12 Cup, White Paper, 8-inch, 50-Count Boxes (Pack of 12) (Packaging May Vary)	$ 20 88	$ 13 46
Mr. Coffee 12-Cup Programmable Coffee Maker, Stainless Steel	$ 33 99	$ 29 99
Mr. Coffee 3-Quart Iced Tea and Coffee Maker, Blue	$ 28 39	$ 28 53
Mr. Coffee PLD12-2 MR. COFFEE 12-Cup Replacement Decanter	$ 19 87	$ 19 91
Hamilton Beach 49976 Flex brew 2-Way Brewer Programmable Coffee Maker, Black	$ 70 54	$ 70 56
Mr. Coffee Café 20-Ounce Steam Automatic Espresso and Cappuccino Machine, Silver/Black	$ 104 95	$ 82 49
Mr. Coffee Single Serve 24 oz. Coffee Brewer, Black	$ 103 80	$ 114 99
Mr. Coffee Café Barista Premium Espresso & Cappuccino System, Silver	$ 155 99	$ 165 70

Slika 2 Primjer korištenja alata Importio Izvor: https://www.g2.com/products/import-io-2017-12-19/reviews , pristupljeno 13.08.2020

5.2. Mozenda

Mozenda je web scraping alat za obradu podataka. Web scraping je sistematizirana ekstrakcija sadržaja (tekstualnog ili medijskog) s web-stranica (Marić 2020.)[4]. Omogućava prikupljanje seta podataka koji su korisniku potrebni te zatim obradu tih podataka i priprema ih za korištenje. Mozenda automatski prikuplja informacije organizirane u popisima na web stranicama koje su prethodno odredili korisnici i omogućuje korisnicima da izrade agente za prikupljanje ovih podataka. Alat može prikupiti podatke iz različitih kategorija i različitih vrsta straničnih struktura. Imena i pridružene vrijednosti se također automatski prepoznaju. Na slici 3 prikazan je primjer obrade podataka u alatu Mozenda.

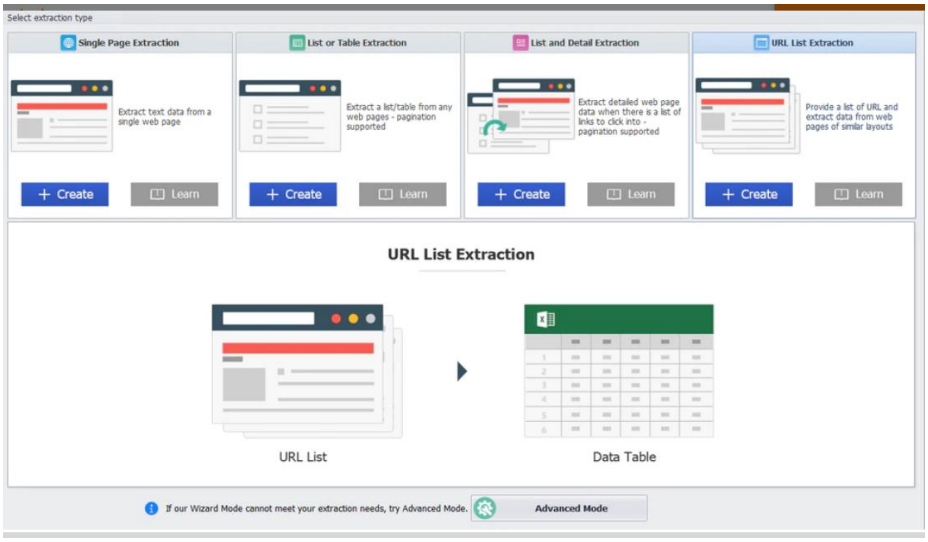
[4] Marić Domagoj,(2020), Kako obogatiti skup podataka podacima s weba s pomoću web scrapinga, https://lider.media/poslovna-scena/hrvatska/kako-obogatiti-skup-podataka-podacima-s-weba-s-pomocu-web-scrapinga-132084 , Pristupljeno 13.08.2020.



Slika 3 Primjer obrade podataka u alatu Mozenda, Izvor https://www.softwareadvice.com/bi/mozenda-profile/ , pristupljeno 3.06.2020.

5.3. Octoparse

Octoparse je besplatan web scraping alat koji pretvara web stranice u strukturirane podatke bez potrebe za pisanjem programskog koda, što je i glavna razika u odnosu na Mozendu. Jednostavan je za korištenje jer radi na sličnom principu kao i web pretraživanja, npr. pretraživanja web stranica i kreiranja korisničkih računa na web aplikacijama. Korisnici mogu kliknuti element na web stranici kako bi odabrali vrstu podataka za izdvajanje. Octoparse omogućava korisnicima da pokreću više zadataka istovremeno. Zadaci se mogu zakazati u redovitim intervalima ili se mogu izvoditi u stvarnom vremenu (Software advice 2020.). Na slici 4 opisan je primjer korištenja alata Octoparse.



Slika 4 Primjer korištenja alata Octoparse Izvor:https://www.softwareadvice.com/bi/octoparse-profile/ ,pristupljeno 3.06.2020.

5.4 NoSQL

NoSQL u prijevodu označava "ne samo SQL". Alternativa je tradicionalnim relacijskim bazama podataka u koje se podaci smještaju u tablice, a shema podataka pažljivo je osmišljena prije nego što se baza podataka izgradi. NoSQL je koristan za rad s velikim količinama distribuiranih podataka. NoSQL omogućava pristup dizajniranju baza podataka koji može primiti široki raspon modela podataka, uključujući ključeve, vrijednosti dokumenata i grafove (Datamanagment, 2017.).

5.4.1.Evolucija NoSQL-a

Berkley DB je bio početni NoSQL baze podataka, a opisan je kao baza podataka koja podržava posebne potrebe pohrane aplikacije. Razvijen je na Sveučilištu Kalifornija početkom 1990-ih.Ovaj softver otvorenog koda pružao je jednostavnu pohranu ključa i vrijednosti. Ostale NoSQL baze podataka koje se ističu uključuju NoSQL baze podataka koje se nalaze u oblaku, poput Amazon DynamoDB, Google BigTable, kao i Apache Cassandra i MongoDB. Osnovne klasifikacije NoSQL baze podataka samo su vodiči. Vremenom su se miješali elementi iz različitih obiteljskih stabala NoSQL baza podataka kako bi postigli korisne sustave. Ponekad se NoSQL elementi miješaju sa SQL elementima, stvarajući različite baze podataka koje se nazivaju multimodelskim bazama podataka. Tipovi baza podataka koji pripadaju u obitelj NoSQL baza prikazane su u tablici 2.

Tablica 2 Popis i karakteristike baza iz obitelji NoSQL

Izvor: https://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL prlagođeno i prevedeno od strane autora, pristupljeno 1.06.2020.

Dokument-baze podataka	Graf baze podataka	Baze podataka ključ-vrijednost	Stupčane baze podataka
Pohranjuje podatke u obliku dokumenata u formatima kao što je JSON.	Naglašava veze između podatkovnih elemenata između „čvorova“ kako bi se ubrzali upiti	Koristi jednostavan model u kojem se uparuje jedinstveni ključ i njegova pripadajuća vrijednost	Također poznata kao tablična baza podataka – pohranjuje podatke u tablicama koje mogu imati velik broj stupaca
Koristi se u svrhe upravljanja sadržajem i kontrole web i mobilnih aplikacija.	Koristi se za sustav preporuka i GIS aplikacije(kartografske aplikacije)	Koristi se za klikovne podatke i prijave u aplikacije	Koristi se pri pretraživanju interneta i Web aplikacija
Primjeri: Couchbase Server, CouchDB, MarkLogic, Mongo DB	Primjeri: Allegrograph, IBM, Graph, Neo4j	Primjeri: Aerospike, DynamoDB, Redis, Riak	Primjeri: Accumulo, Cassandra, Hbase, Hypertable, SimpleDB

Baze podataka ključ-vrijednost (key-value databases)

Baze podataka ključ-vrijednost implementiraju jednostavan model podataka koji spaja jedinstveni ključ s pridruženom vrijednošću. Budući da je ovaj model jednostavan, može dovesti do razvoja baza podataka s ključnom vrijednošću, koje su izuzetno uspješne i vrlo skalabilne za upravljanje sesijama i predmemoriranje u web aplikacijama. Implementacije se razlikuju u načinu na koji su orijentirane za rad s RAM-om, solid-state diskovima ili diskovnim pogonima. Primjeri uključuju Aerospike, Berkeley DB, MemchacheDB, Redis i Riak.

Baze podataka dokumenata (Document database)

Baze podataka dokumenata, koje se nazivaju i pohranjivanjem dokumenata, pohranjuju polustrukturirane podatke i opise tih podataka u formatu dokumenta. Oni omogućuju programerima da stvaraju i ažuriraju programe bez potrebe za referentnom shemom. Upotreba baza podataka dokumenata povećala se zajedno s upotrebom JavaScripta i JavaScript Object Notation (JSON), formata razmjene podataka koji je stekao veliku važnost među programerima web aplikacija, mada se mogu koristiti i XML i drugi formati podataka. Baze podataka

dokumenta koriste se za upravljanje sadržajem i rukovanje podacima mobilnih aplikacija. Couchbase Server, CouchDB, DocumentDB, MarkLogic i MongoDB su primjeri baza podataka (Datamanagment, 2017.).

Stupčane baze podataka(Wide-column stores)

Stupčaste baze podataka organiziraju podatake kao stupce, a ne kao redove. Mogu se naći i u bazama podataka SQL i NoSQL.Brzo pretražuju velike količine podataka u odnosu na uobičajene relacijske baze podataka. Spremnik podataka sa širokim stupcima(wide-column stores) može se koristiti za mehanizme preporuka, kataloge, otkrivanje prijevara i druge vrste obrade podataka. Google BigTable, Cassandra i HBase primjeri su tabličnih stupčastih baza podataka (Datamanagment, 2017.).

Grafičke baze podataka(Graph database)

Grafičke baze podataka organiziraju podatke kao čvorove, koji su poput zapisa u relacijskoj bazi podataka i rubova, koji predstavljaju veze između čvorova. Budući da grafički sustav pohranjuje odnos između čvorova, može podržati predstavljanje odnosa podataka. Također, za razliku od relacijskih modela koji se oslanjaju na stroge sheme, grafički model podataka može se razvijati tijekom vremena i upotrebe. Grafičke baze podataka primjenjuju se u sustavima koji moraju mapirati odnose, poput rezervacijskih sustava ili upravljanja odnosima s kupcima. Primjeri baza podataka grafova uključuju AllegroGraph, IBM Graph, Neo4j i Titan.

Postoji velik broj NoSQL baza podataka za odabir, stoga se postavlja pitanje; Kako odrediti koja je najbolja za određeni skup podataka? Pronalaženje prave NoSQL baze podataka ostaje izazov jer postoji mnogo opcija s različitim arhitekturama i slučajevima primarne uporabe. Mnoge organizacije smatraju da su njihove potrebe za podacima zadovoljene relacijskim bazama podataka, NoSQL se i dalje smatra standardom za određene slučajeve upotrebe, uključujući IoT upravljanje podacima, društvene medije i analitiku u stvarnom vremenu, gdje velika količina podataka dolazi brzo i često nestrukturirana. Prema nedavnom istraživanju kompanije ScaleGrid, davatelja baze podataka kao usluge, preko 75% više korisnika baze podataka uključuje i SQL i NoSQL kako bi upravljali svojim podacima (Datamanagment, 2017.).

Multi-model i NoSQL

Iako postoji sve veća potreba za više modelnim bazama podataka, još uvijek nije uobičajena u najpopularnijim NoSQL bazama podataka. Mnoge organizacije koriste više baza podataka, neke podijele svoje podatke između relacijskih i NoSQL baza podataka, za različite modele podataka. Sigurnost je jedan od problema za NoSQL baze podataka, a jedna od prednosti većine modela je mogućnost da više analitičara ima pristup dostupnim podacima. Posljednjih godina NoSQL modeli postavili su jače mjere sigurnosti, od automatske enkripcije podataka do provjere autentičnosti korisnika i metoda autorizacije. Uz to, kako bi se poboljšala sigurnost, mnoge baze podataka NoSQL uključuju standardne evidencije revizije kako bi se rano upozorilo poduzeća na nedosljednosti u podacima ili zapisnicima, tako da poduzeća mogu i sama biti aktivna u zaustavljanju kršenja podataka.[5]

5.4.2. Usporedba Sql i NoSQL baza podataka

NoSQL („ne SQL“ ili „ne samo SQL“) baze podataka razvijene su krajem 2000-ih s naglaskom na skalabilnost, brze upite, što omogućuje česte promjene aplikacija i pojednostavljenje programiranja za programere (MongoDB.com, 2020..). Ključna razlika je u strukturiranosti podataka. SQL baza podataka je sklonija strukturiranim unaprijed određenim podatcima i oblicima dok je NoSQL baza podataka prilagođena nestrukturiranim podatcima. Tablica 3 pobliže prikazuje razlike između SQL i NoSQL baza podataka.

[5] Karla E.Joyce, veljača 2020, NoSQL database comparison to help you choose the right store, https://searchdatamanagement.techtarget.com/infographic/NoSQL-database-comparison-to-help-you-choose-the-right-store, pristupljeno 1.06.2020..

Tablica 3 Usporedba SQL I NoSQL baze podataka

Izvor: https://shishirkumarblog.wordpress.com/technical/sql-vs-nosql-the-cap-theorem/ , prilagođeno i prevedeno od strane autora, pristupljeno 14.07.2020.

SQL	NoSQL
SHEMA: Nepromijenjiva – Tablice i Stupci su unaprijed određeni	**SHEMA: Promijenjiva shema – Može pohraniti entitet i njegov atribut koji nisu unaprijed određeni**
POHRANA: Tablica (Redak -> Entitet, Stupac -> Atribut) **RBMS: Oracel, IBM DB2, Microsoft SQL Server, MySQL**	**POHRANA** • **Ključ/vrijednost-Redis, Dynamo** • **Document – MongoDB** • **Graph – Neo4j, InfiniteGraph** • **Wide-columne – Cassandra, HBase**
UPIT: SQL	**UPIT: UnSQL**
SKALABILNOST: Prikladno za vertikalno stupnjevanje	**SKALABILNOST: Prikladno za Horizontalno stupnjevanje**
ACID(Atomicity, Consistency, Isolation and Durability) usklađenost	**Većina NOSql baza podataka podržava ACID svojstva**

6. Hadoop - opis, svojstva, pristup i način funkcioniranja

Projekt Apache Hadoop razvija open-source softver za pouzdano, skalabilno i distribuirano računanje. Biblioteka softvera Apache Hadoop okvir je koji omogućuje distribuiranu obradu velikih skupova podataka preko klastera računala pomoću jednostavnih modela programiranja (Hadoop.org, 2020.). Hadoop su kreirali računalni znanstvenici Doug Cutting i Mike Cafarella, koji su u početku podržavali obradu u pretraživaču s otvorenim kodom Nutch i web pretraživaču. Nakon što je Google objavio tehničke radove u kojima je detaljno opisao svoj Google File System i programski okvir MapReduce u 2003. i 2004., Cutting i Cafarella izmijenili su ranije tehnološke planove i razvili JavaR zasnovanu implementaciju MapReduce i datotečni sustav po uzoru na Googleov.[6] Dizajniran je tako da se ne oslanja na hardver za isporučivanje velike dostupnosti, sama je biblioteka dizajnirana za otkrivanje i rukovanje kvarovima na aplikacijskom sloju, tako da pruža visoko dostupnu uslugu na vrhu klastera računala, od kojih je svako sklono kvarovima. Glavne karakteristike Hadoop sustava su: brza pohrana podataka, računalna snaga, tolerancija na kvarove, fleksibilnost, skalabilnost i niska cijena.

Sposobnost brzog pohranjivanja i obrade ogromne količine bilo koje vrste podataka

[6] Margot Rouse, (2019), Hadoop, https://searchdatamanagement.techtarget.com/definition/Hadoop , pristupljeno 3.06.2020.

Budući da se količina podataka i raznolikost stalno povećava, posebno s društvenih medija i Interneta stvari (IoT) ključno pitanje je računalna snaga. Računalni model ima sposobnost brze obrade velikih podataka. Što se više računalnih čvorova koristiti, postoji i veća sposobnost obrade podataka (Sas.com, 2020.).

Tolerancija na kvarove

Obrada podataka i sama aplikacija zaštićena je od kvara hardvera. Ako čvor propadne, zadatci se automatski preusmjeravaju na druge čvorove kako bi se osigurao neometan rad sustava. Višestruke kopije svih podataka automatski se pohranjuju.

Fleksibilnost

Za razliku od tradicionalnih relacijskih baza podataka, podatci se ne moraju prethodno obraditi prije njihovog pohranjivanja odnosno ne moraju biti unaprijed definirani. Moguće je pohraniti bilo koju količinu podataka te kasnije odlučiti koja će biti svrha tih podataka. To uključuje nestrukturirane podatke bilo koje vrste.

Niska cijena

Hadoop sustav omogućava nisku cijenu pohrane podataka. Okvir otvorenog koda je besplatan i koristi hardver za pohranu velikih količina podataka.

Skalabilnost

Sustav može lako rasti i obrađivali više podataka jednostavnim dodavanjem čvorova. Također postoje i određeni izazovi korištenja Hadoop-a. MapReduce programiranje ne odgovara na sve probleme. Dobar je za jednostavne informacije i probleme koji se mogu podijeliti u neovisne jedinice, ali nije učinkovit za iterativne i interaktivne analitičke zadatke. MapReduce zahtijeva veliku datoteku, budući da čvorovi međusobno ne komuniciraju osim putem sortiranja i mješanja, iterativni algoritmi zahtijevaju da se završe višestruke faze izmjene. MapeReduce odnosi se na model programiranja i pripadnu implementaciju za obradu i generiranje velikih podataka na klasteru računala. Temelji se na principu paralelnog računanja nad nekim skupom podataka. Drugi je izazov usredotočen na pitanje sigurnosti podataka, iako se pojavljuju novi alati i tehnologije. Kerberos protokol provjere autentičnosti odličan je korak za osiguravanju Hadoop okruženja. Slijedeći nedostatak predstavlja to što Hadoop nema jednostavne alate za

upravljanje podacima, čišćenje i upravljanje metapodacima s punim značajkama. Posebno nedostaju alati za kvalitetu podataka i standardizaciju.[7]

6 .1 Komponente Hadoop sustava i način funkcioniranja

Glavne komponente u prvoj inačici Hadoopa bile su MapReduce, HDFS (Hadoop Distributed File System) i Hadoop Common, skup zajedničkih uslužnih programa i biblioteke. MapReduce koristi mape i reducira funkcije da bi podijelio zadatke za obradu u više zadataka koji se izvode na čvorovima klastera u kojima se pohranjuju podaci, a zatim da kombinira ono što zadaci proizvode u koherentan skup rezultata. MapReduce je u početku funkcionirao kao Hadoopov procesor i menadžer resursa klastera, koji je HDFS izravno vezao za njega i ograničio korisnike u pokretanju batch aplikacija MapReduce. To se promijenilo u Hadoopu 2.0, koji je postao dostupan u listopadu 2013. kada je objavljena verzija 2.2.0. Predstavio je Apache Hadoop YARN, novo upravljanje resursima klastera i tehnologiju zakazivanja poslova koji su preuzeli te funkcije od MapReducea. YARN (Yet Another Resource Negotiator ili još jedan pregovarač o resursima). Time se smanjilo oslanjanje na MapReduce te je Hadoop postao dostupan drugim procesorskim sustavima i raznim aplikacijama. Na primjer, Hadoop sada može pokretati aplikacije na sustavima Apache Spark, Apache Flink, Apache Kafka i Apache Storm. U klasterima se Hadoop, YARN nalazi između HDFS-a i procesnih sustava koje implementiraju korisnici. Upravitelj resursa koristi kombinaciju spremnika, koordinatora aplikacija i nadzornih sredstava na čvoru radi dinamičke raspodjele resursa klastera aplikacijama i nadzora nad izvršavanjem poslova obrade u decentraliziranom procesu. YARN podržava više pristupa raspoređivanju poslova i nekoliko metoda za planiranje poslova na temelju dodijeljenih resursa klastera. Hadoop 3.0.0 je sljedeća velika verzija Hadoopa. Apache je objavio u prosincu 2017., dodao je značajku YARN federacije namijenjenu YARN-u da podrži desetine tisuća čvorova ili više u jednom klasteru, što je u odnosu na prethodni limit od 10.000 čvorova. Nova verzija također je uključivala podršku za GPU i kodiranje brisanja, alternativu kopiranju podataka za koji je potrebno znatno manje prostora za pohranu. Naknadna ažuriranja 3.1.x i 3.2.x omogućila su korisnicima Hadoop-a da pokreću YARN kontejnere unutar Docker-ovih i uvela YARN servisni okvir koji funkcionira kao platforma za orkestraciju spremnika. Dvije su nove Hadoop komponente dodane s tim izdanjima: stroj za strojno učenje pod nazivom Hadoop

[7] Hadoop What it is and why it matters, 2020, https://www.sas.com/en_us/insights/big-data/hadoop.html#hadoopusers , pristupljeno 3.06.2020

Submarine i objektna trgovina Hadoop Ozone, koji je izgrađen na bloku za pohranu podatak Data Store i dizajniran je za upotrebu u lokalnim sustavima. Tablica 4.

Tablica 4 Glavne sastavnice Hadoop sustava

Izvor: https://searchdatamanagement.techtarget.com/definition/Hadoop ,prilagođeno i prevedeno od strane autorice M.Rouse,pristupljeno 14.08.2020.

HDFS (Hadoop Distributed File System)	YARN	MapReduce	Hadoop Common
Sustav dokumenata koji upravlja skladištem i pristupom podataka s različitih čvorova Hadoop klastera. Hadoop klaster je zaslužan za alociranje resursa sustava aplikacijama i zakazivanju zadataka	**Hadoop klaster je zaslužan za alociranje resursa sustava aplikacijama i zakazivanju zadataka**	**Programski framework i procesor koristi se za pokretanje large – scale aplikacija Hadoop sustava uslužnih programa i datoteka**	**Sustav uslužnih programa i biblioteka koje omogućuju osnovne mogućnosti na zahtjev drugih dijelova Hadoop sustava**

7. Primjena nestrukturiranih podataka

Tvrtke koriste podatke prikupljene u svojim sustavima za poboljšanje poslovanja, pružanje bolje usluge kupcima, stvaranje personaliziranih marketinških kampanja temeljenih na specifičnim preferencijama kupaca i u konačnici, povećavanju profitabilnosti. Tvrtke koje koriste nestrukturirane podatke imaju potencijalnu konkurentsku prednost u odnosu na one koja to ne čine, budući da mogu lakše donositi poslovne odluke zbog više informacija o konkretnom tržištu, pod uvjetom da te podatke učinkovito koriste. Na primjer velike baze podataka mogu tvrtkama pomoći u kreiranju odgovarajuće marketinške strategije jer će time imati više podataka o potencijalnim kupcima odnosno klijentima. Veliki podaci dolaze iz bezbroj različitih izvora, kao što su sustavi poslovnih transakcija, baze podataka o kupcima, medicinska evidencija, mobilne aplikacije, društvene mreže, baze znanstvenih istraživanja, strojno generirani podaci i senzori podataka u stvarnom vremenu koji se koriste u internetu stvari (IoT) okruženja. Podaci se mogu ostaviti u svom neobrađenom obliku u velikim podatkovnim sustavima ili unaprijed obraditi pomoću alata za vađenje podataka ili softvera za pripremu podataka tako da su spremni za određenu upotrebu analitike. Nestrukturirani podatci se mogu iskorištavati u različite svrhe, a ovo su neke od njih:

Analiza ponašanja korisnika

Uključuje ispitivanje pokazatelja ponašanja korisnika i promatranje angažmana kupaca u stvarnom vremenu kako bi se uspoređivali proizvodi, usluge i ovlasti robne marke jedne tvrtke s konkurencijom.

Slušanje na društvenim medijima

Informacije o tome što ljudi govore na društvenim mrežama o određenom poslu ili proizvodu koji nadilazi ono što se može istražiti anketom . Ovi se podaci mogu koristiti za prepoznavanje ciljane publike za marketinške kampanje promatranjem aktivnosti koja se bavi određenim temama iz različitih izvora.

Marketing analiza

Uključuje informacije koje se mogu koristiti za više informiranosti i inovativnosti promocije novih proizvoda, usluga i inicijativa.

Analiza zadovoljstva kupaca

Sve prikupljene informacije mogu otkriti kako se kupci osjećaju u vezi s tvrtkom ili markom, mogu li se pojaviti bilo kakvi potencijalni problemi, kako se može sačuvati lojalnost kupaca i kako se mogu poboljšati usluge poslovnog subjekta.

Osim u poslovanju nestrurkturirani podatci i velike skupine podataka (big data) mogu se koristiti i u javnom sektoru te različitim granama gospodarstva. Medicinski istraživači također koriste velike baze podataka (big data) za utvrđivanje čimbenika rizika od bolesti za pomoć u dijagnosticiranju bolesti i stanja u pojedinih bolesnika. Uz to, podaci dobiveni iz elektroničkih zdravstvenih kartona, društvenih medija, interneta i drugih izvora pružaju zdravstvenim organizacijama i vladinim agencijama najnovije informacije o prijetnjama ili epidemijama zaraznih bolesti. U energetskoj industriji, velike baze podataka pomažu naftnim i plinskim kompanijama da identificiraju potencijalna mjesta bušenja i nadgledaju rad cjevovoda; slično, komunalni sustavi ga koriste za praćenje rada električnih mreža. Tvrtke za financijske usluge koriste velike podatkovne sustave za upravljanje rizikom i analizu tržišnih podataka u stvarnom vremenu. Prijevozničke tvrtke oslanjaju se na velike podatke kako bi upravljali svojim lancima opskrbe i optimizirali rute isporuke. Ostale upotrebe vlade uključuju reagiranje u kriznim situacijama, sprečavanje kriminala i pametne gradske inicijative. Konačno, vrijednost i učinkovitost velikih podataka ovise o stručnjacima koji imaju zadatak da razumiju podatke i

formuliraju odgovarajuće upite za usmjeravanje projekata analize velikih podataka. Neki alati za velike podatke zadovoljavaju specijalizirane niše i omogućavaju korisnicima da koriste svakodnevne poslovne podatke u aplikacijama za prediktivnu analizu. IBM SPSS Modeler softverski je alat za prediktivnu analitiku. Korisnicima omogućuje kreiranje prediktivne analitike otkrivajući uzorke i odnose među podacima. Podaci mogu biti strukturirani ili nestrukturirani (I.S, 2014.). Korisnicima omogućuje analitiku i pretraživanje podataka bez potrebe za pisanjem programskog koda. Ostale tehnologije, poput uređaja sa velikim podacima koji se temelje na Hadoopu, pomažu tvrtkama implementirati odgovarajuću računalnu infrastrukturu za rješavanje velikih podataka, istovremeno smanjujući potrebu za hardverom i distribuiranim softverskim znanjem.

7.1 Primjer korištenja Hadoop sustava

Hadop sustav je promijeniv i fleksibilan te mnogim tvrtkama odgovara to što lako mogu ažurirati svoj podatkovni sustav i prilagoditi situaciji u kojoj se nalaze. Iz tog razloga je korištenje Hadoop sustava široko rasprostranjeno. Također mnoge velike tvrtke koriste upravo Hadoop sustav kao svoje skladište podataka upravo zato što je to jeftin sustav koji se može ažurirati. Hadoop sustav je posebno koristan u financijskom sektoru. Financijski sektor teži ka digitalizaciji svojih usluga i tehnološkom napretku. Dvije trećine organizacija u financijskom sektoru kažu da im Hadoop pomaže povećati poslovnu agilnost i operativnu učinkovitost (Wilson, 2017.). S obzirom da banke i druge financijske ustanove posjeduju velike količine podataka od velike važnosti za zaposlenike i korisnike, Hadoop sustav olakšava obradu i pohranu tih podataka. Hadoop se često koristi u pružanju financijskih usluga zbog svoje moći i u obradi podataka i u analizi (N.Ismail,2017.). Stoga ovaj sustav pomaže u nastanku sustava za procjenu kreditnog rizika, investicijskim modelima, sustavima za prikazivanje trendova. Hadoop sustav također može pomoći bankama u otkrivanju i otklanjanju potencijalnih prevara. To je moguće pomoću big data tehnologiije odnosno tehnologije zasnovane na velikim bazama podataka. Analitikom takvih podataka banke mogu doći do korisnih informacija o neobičnim ponašanjima vezanim za kreditne kartice i tako predvidjeti krađe. Također putem kvalitetne analize podataka banke mogu poboljšati odnos sa svojim klijentima jer će time imati veći broj informacija o klijentima i moći će im ponuditi odgovarajuće usluge. Sva poduzeća, posebno financijske tvrtke, trebaju sada koristiti velike podatke i Hadoop tehnologije kako bi ostvarila svoj najveći potencijal, posebno s velikom količinom podataka i transakcija koje se svakodnevno objedinjuju (M. Nemschoff, 2014.).

Zaključak

U današnje vrijeme brzog tehnološkog razvoja raširenosti korištenja interneta te internetskih i mobilnih aplikacija povećava se i količina podataka. Određeni dio podataka je lako dostupan svima za korištenje i pretraživanje. Međutim dio tih podataka nalazi se u strukturiranom obliku,a drugi dio se nalazi u nestrukturiranom obliku. Nestrukturirane podatke je teško pretraživati,obrađivati i koristiti u svrhe poslovanja ili osobne svrhe. Postoje razni alati te baze podataka kao što su npr: Mozenda, Octoparse i Importio koji su u radu pobliže opisani. Također su opisane i NoSQl baze podataka te Hadoop sustav kao i prednosti i mane takvih sustava. Međutim ne postoji jedinstveno rješenje koje bi pomoglo pri riješavanju problema pojave velikog broja nestrukturiranih baza podataka. Iz tog razloga svaki poslovni subjekt ili korisnik mora pronaći najbolji način za vlastito pretraživanje nestrukturiranih podataka. Postoje i razne primjene ovakvih sustava u gospodarskom i javnom sektoru što još jednom dokazuje važnost razumijevanja rada ovakvih sustava. Nestrukturirani podatci mogu biti od velike važnosti za korisnika, organizaciju i poduzeće i ukoliko se takvi podatci ne obrađuju i ne pohranjuju, organizacije gube velike količine podataka koji su mogli biti potencijalno korisni za njihov budući rad.

Literatura

1. Upravljanje dokumentima http://www.senso-is.hr/upravljanje-dokumentima.aspx (pristupljeno 5.svibnja.2020)
2. https://www.datamation.com/big-data/structured-vs-unstructured-data.htm (pristupljeno 5.svibnja.2020)
3. There's no such thing as unstructured data http://analytics-magazine.org/theres-no-such-thing-as-unstructured-data/ (pristupljeno 25.svibnja.2020)
4. Import.io -Mission-critical web data https://www.import.io/ (pristupljeno 26.svibnja.2020)
5. Import.io Reviews & Product Details https://www.g2.com/products/import-io-2017-12-19/reviews (pristupljeno 13.kolovoza.2020)
6. NoSQL (Not Only SQL database) https://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL (pristupljeno 1.lipnja.2020)
7. NoSQL database comparison to help you choose the right store https://searchdatamanagement.techtarget.com/infographic/NoSQL-database-comparison-to-help-you-choose-the-right-store (pristupljeno 1.lipnja.2020)
8. https://hadoop.apache.org/ (pristupljeno 3.lipnja.2020)
9. Hadoop-What it is and why it matters https://www.sas.com/en_us/insights/big-data/hadoop.html#close (pristupljeno 3.lipnja.2020)
10. Hadoop https://searchdatamanagement.techtarget.com/definition/Hadoop (pristupljeno 3.lipnja.2020)
11. Mozenda https://www.mozenda.com/ (pristupljeno 3.lipnja.2020)
12. What is Mozenda? https://www.softwareadvice.com/bi/mozenda-profile/ (pristupljeno 3.lipnja.2020)
13. Octoparse Software https://www.softwareadvice.com/bi/octoparse-profile/ (pristupljeno 3. lipnja.2020)
14. What is Octoparse? https://www.softwareadvice.com/bi/octoparse-profile/ (pristupljeno 3.lipnja.2020)
15. Big Data https://searchdatamanagement.techtarget.com/definition/big-data (pristupljeno 21.srpnja.2020)
16. 7 Big Data Examples: Applications of Big Data in Real Life https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/ (pristupljeno 21.srpnja.2020)
17. Ten Steps for Analyzing Unstructured Data https://medium.com/@vratulmittal/10-steps-for-analyzing-unstructured-data-1b4f48544c9a (pristupljeno 22.srpnja.2020)
18. Kako obogatiti skup podataka podacima s weba s pomoću web scrapinga https://lider.media/poslovna-scena/hrvatska/kako-obogatiti-skup-podataka-podacima-s-weba-s-pomocu-web-scrapinga-132084 (pristupljeno 13.kolovoza.2020)
19. What is Enterprise Content Management? https://www.laserfiche.com/ecmblog/what-is-enterprise-content-management-ecm/ (pristupljeno 14.kolovoza.2020)

20. Who’s Using Hadoop? And What are They Using It For? https://blog.syncsort.com/2015/06/big-data/whos-using-hadoop-and-what-are-they-using-it-for/ (pristupljeno 15.kolovoza.2020)
21. Hadoop in finance: big data in the pursuit of big bucks https://www.information-age.com/hadoop-finance-big-data-pursuit-big-bucks-123465206/ (pristupljeno 15.kolovoza.2020)