Analiza velikih skupova podataka u oblaku računala

Krpan, Biljana

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
ELEKTROTEHNIČKI FAKULTET

Sveučilišni studij

ANALIZA VELIKIH SKUPOVA PODATAKA U OBLAKU RAČUNALA

Diplomski rad

Biljana Krpan

Osijek, 2016.

ELEKTROTEHNIČKI FAKULTET OSIJEK
Sveučilište Josipa Jurja Strossmayera u Osijeku

Obrazac D1: Obrazac za imenovanje Povjerenstva za obranu diplomskog rada

Osijek,

Odboru za završne i diplomske ispite

Imenovanje Povjerenstva za obranu diplomskog rada

Ime i prezime studenta:	Biljana Krpan
Studij, smjer:	Sveučilišni diplomski studij, smjer Procesno računarstvo
Mat. br. studenta, godina upisa:	D-585R, 2013. godina
Mentor:	prof.dr.sc. Goran Martinović
Sumentor:	
Predsjednik Povjerenstva:	
Član Povjerenstva:	
Naslov diplomskog rada:	Analiza velikih skupova podataka u oblaku računala
Primarna znanstvena grana rada:	**Računarstvo**
Sekundarna znanstvena grana (ili polje) rada:	
Zadatak diplomskog rada:	U diplomskom radu treba proučiti zahtjeve, načine i alate za analizu velikih skupova podataka u oblaku računala. Na testnom primjeru velikog skupa podataka treba primijeniti najpovoljniju kombinaciju postupaka i alata, te prikladno analizirati dobivene rezultate s gledišta perfomansi i primjenjivosti.
Prijedlog ocjene pismenog dijela ispita (diplomskog rada):	
Kratko obrazloženje ocjene prema Kriterijima za ocjenjivanje završnih i diplomskih radova:	Primjena znanja stečenih na fakultetu: Postignuti rezultati u odnosu na složenost zadatka: Jasnoća pismenog izražavanja: Razina samostalnosti:

Potpis sumentora:	Potpis mentora:

Dostaviti:

1. Studentska služba

U Osijeku, godine Potpis predsjednika Odbora:

ETFOS
ELEKTROTEHNIČKI FAKULTET OSIJEK
Sveučilište Josipa Jurja Strossmayera u Osijeku

IZJAVA O ORIGINALNOSTI RADA

Osijek,

Ime i prezime studenta:	Biljana Krpan
Studij :	Sveučilišni diplomski studij, smjer Procesno računarstvo
Mat. br. studenta, godina upisa:	D-585R, 2013. godina

Ovom izjavom izjavljujem da je rad pod nazivom**:**

Analiza velikih skupova podataka u oblaku računala

izrađen pod vodstvom mentora

prof.dr.sc. Goran Martinović

i sumentora

moj vlastiti rad i prema mom najboljem znanju ne sadrži prethodno objavljene ili neobjavljene pisane materijale drugih osoba, osim onih koji su izričito priznati navođenjem literature i drugih izvora informacija.
Izjavljujem da je intelektualni sadržaj navedenog rada proizvod mog vlastitog rada, osim u onom dijelu za koji mi je bila potrebna pomoć mentora, sumentora i drugih osoba, a što je izričito navedeno u radu.

Potpis studenta:

SADRŽAJ

1. UVOD

Razvoj modernih informacijskih i komunikacijskih tehnologija omogućio je prikupljanje velikih skupova podataka. Veliki skupovi podataka su skupovi podataka koje je vrlo teško obraditi tradicionalnim postupcima i programskim rješenjima. Tri temeljne odrednice velikih podataka su obujam, raznolikost i brzina njihova nastanka. Obujam se odnosi na količinu podataka koja neprestano raste, dok se pod brzinom misli na rastuću frekvenciju pojavljivanja podataka. Ono što velike podatke čini zahtjevnijima za analizu jest raznovrsnost njihovih izvora. Razvojem tehnologija povećao se broj web izvora kao što su društvene mreže, mobilni telefoni i njihove aplikacije, digitalna televizija, razni senzorski podaci Internet objekata i sl.

Neorganizirani i nekategorizirani podaci često nemaju vidljivu primjenu. Uz rad i uloženi kapital, precizna informacija je čimbenik od velike važnosti koji će odrediti hoće li neki poslovni pothvat završiti uspješno ili ne. Kod velikih skupova podataka, ključne su analiza i pravilna interpretacija koje se odvijaju na najnovijim platformama. Računarstvo u oblaku je infrastruktura koja omogućuje pristup računalima i njihovim funkcionalnostima putem Interneta. Implementacijom računarstva u oblaku, korisnici oblaka (podatkovni znanstvenici i drugi) putem različitih mrežnih usluga, puno lakše pristupaju problemu pohrane i obrade velikih skupova podataka. Analizom velikih skupova podataka u oblaku, mogu se dobiti razne povratne informacije korisne u svim sektorima, od zdravstva, financija, industrije, poduzetništva do znanstvenih i obrazovnih institucija.

Cilj ovog rada je istražiti zahtjeve obrade i tehnologije velikih skupova podataka u oblaku računala te primjenom alata *Pig, Hive, HCatalog i MS Excel* na otvorenoj datoteci zapisa s internet stranica, izvršiti analizu i prikazati neke od mogućih primjena u poslovanju.

Drugo poglavlje pobliže predstavlja temeljnu definiciju velikih skupova podataka, dimenzije i ključne karakteristike, povijest razvoja te utjecaj na poslovanje. Treće poglavlje opisuje temeljnu strukturu ekosustava *Hadoop* koji se koristi pri pohrani, obradi i analizi velikih skupova podataka. U istom dijelu predstavljena je podatkovna platforma *Hortonworks* i njena struktura te način instalacije i programiranja unutar platforme. Pokazni primjeri opisani su u četvrtom poglavlju rada gdje su kroz stvarne zahtjeve neke tvrtke prikazane manipulacijske moći nad velikim podacima te njihova primjena u poslovnom životu.

1.1. Zadatak diplomskog rada

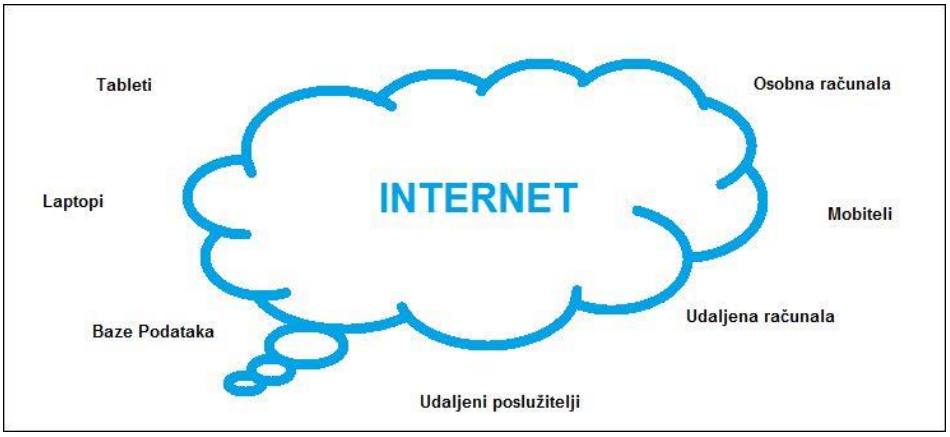
Zadatak ovog diplomskog rada je proučiti zahtjeve, načine i alate za analizu velikih skupova podataka u oblaku računala. Na testnom primjeru velikog skupa podataka treba primijeniti najpovoljniju kombinaciju postupaka i alata te prikladno analizirati dobivene rezultate s gledišta perfomansi i primjenjivosti.

2. OBLAK RAČUNALA I VELIKI SKUPOVI PODATAKA

Oblak računala ili računarstvo u oblaku (*engl. cloud computing*) prema definiciji NIST-a (*National Institute of Standards and Technology*) je model koji omogućuje jednostavan i „na-zahtjev“ dostupan (*engl. on-demand*) pristup na mrežu dijeljene grupe prilagodljivih računalnih resursa [1]. Takvi računalni resursi mogu se uspostavljati velikom brzinom i pokretati uz minimalan napor za upravljanje ili interakciju s pružateljem usluga (npr. mreže, poslužitelji, podatkovni prostor, aplikacije, usluge, itd.). Računarstvo u oblaku predstavlja „plati-koliko-koristiš“ (*engl. pay-per-use*) model koji omogućuje jednostavan pristup grupama računalnih resursa preko Interneta.

2.1. Osnove računarstva u oblaku

Pojam računarstva u oblaku obuhvaća korištenje mreže udaljenih poslužitelja (umjesto lokalnih poslužitelja) za pohranu, upravljanje i obradu podataka. Budući da mjesta na kojima su udaljeni poslužitelji smješteni i gdje izvršavaju aplikacije i pohranjuju podatke nisu točno definirana, koristi se izraz „u oblaku“. Oblak se, kao što je prikazano na s*lici 2.1,* isto tako vrlo često koristi kao metafora za Internet.



Slika 2.1. Slikovna interpretacija računarstva u oblaku

Računarstvo u oblaku je možda najlakše objasniti na primjeru električne utičnice. Pri korištenju električnih uređaja, ljudi ni ne razmišljaju što se događa iza utičnice niti kako je taj sustav izgrađen i kako funkcionira. On je prisutan u svakom trenutku i naplativ po potrošnji. Isto tako, računarstvo u oblaku možemo zamisliti kao veliki skup računalnih resursa koji su dostupni kada se krajnji korisnik „priključi na oblak“ te iskoristi resurse koji su mu potrebni i plati onoliko resursa koliko je potrošio. Velika prednost pri korištenju oblaka je da korisnici iznajmljuju

infrastrukturu. To znači da više ne postoje troškovi nabave sklopovlja (ne kupuju se poslužitelji, napajanja), niti programski troškovi raznih licenci. Koristi se samo ono što korisnik zahtjeva, a plaća se samo ono što korisnik iskoristi. Upravo iz tih razloga računarstvo u oblaku predstavlja veliki korak u napretku IT evolucije, jer mijenja način na koji razvijamo, implementiramo, održavamo te plaćamo aplikacije i infrastrukturu na kojoj su pokrenute.

2.1.1. Osnovne karakteristike oblaka računala

Stotine milijuna korisnika diljem svijeta koriste usluge koje su bazirane na oblaku, a neke od najpoznatijih su *Gmail, Facebook, Twitter* i drugi. Računarstvo u oblaku *(engl. cloud computing)* je model koji promovira dostupnost i sastoji se od sljedećih ključnih karakteristika:

- **Usluge na zahtjev** (*engl. on-demand self-service*) - korisnik može samostalno odabrati i pokrenuti mogućnosti računalnih resursa kao što su vrijeme poslužitelja i mrežni prostor za pohranu podataka bez interakcije s pružateljem usluga
- **Isporuke usluga preko mreže** (*engl. broad network access*) - isporuka usluga se najčešće odvija preko Interneta
- **Udruživanja resursa** (*engl. resource pooling*) - računalni resursi pružatelja usluga spajaju se kako bi poslužili sve korisnike koristeći višekorisnički model (*engl. multi-tenant model)*
- **Brze elastičnosti** (*engl. rapid elasticity*) – mogućnosti koje pruža oblak računala krajnjem korisniku izgledaju bez ograničenja i mogu se kupiti u bilo kojoj veličini u bilo koje vrijeme
- **Neovisnosti uređaja od mjesta resursa** (*engl. location independent resource pooling*) - omogućava korisnicima pristup sustavu koristeći web preglednik bez obzira na lokaciju i uređaj kojim se pristupa (računalo, mobilni telefon).

2.2. Veliki skupovi podataka

Priča o tome kako su podaci postali „veliki“ počinje još puno prije trenutne tehnološke „buke“ o velikim skupovima podataka. Prema *Oxford English Dictionary-ju* taj je pojam prvi puta korišten 1941. godine [2]. Svako novo spominjanje pojma veliki skupovi podataka (VSP) ili veliki podaci *(engl. big data)* predstavljalo je prekretnicu u povijesti dimenzioniranja količine podataka i prvijence u evoluciji ideje velikih skupova podataka.

2.2.1. Povijesni presjek *VSP-a*

Derek Price, 1961. godine, objavljuje knjigu [3] u kojoj grafički prikazuje znanje znanstvenika promatrajući samo broj objavljenih znanstvenih časopisa i članaka. Pri tome zaključuje da broj novih časopisa/članaka raste eksponencijalno, a ne linearno i da se rast udvostručuje svakih 15 godina u razdoblju od pola stoljeća. Price to naziva „Zakonom eksponencijalnog rasta".

U znanstvenom radu [4] iz 1986. godine, Hal B. Becker procjenjuje da će do 2000. godine poluvodiči s izravnim pristupom memorije spremati 1.25×10^{11} bajtova po kubnom inču.

Prva sveobuhvatna studija izračunavanja opsega novih i originalnih podataka nastalih u svijetu godišnje [5] (ne računajući kopije) objavili su Peter Lyman i Hal R. Varian 2000. godine. U njoj je objavljeno kako se samo u 1999. godini proizvelo 1.5 EB (eksabajta) jedinstvenih informacija, odnosno oko 250 MB (megabajta) za svakog muškarca, ženu i dijete koje živi na Zemlji. Također, smatra se da je ogromna količina jedinstvenih informacija kreirana i pohranjena od strane individualaca (što se kasnije naziva demokracija podataka) te da digitalni podaci nisu samo najveći u ukupnoj proizvodnji nego su ujedno i najbrže rastući. Lyman i Varian ističu „dominaciju digitalnoga" i tvrde da je čak i danas, većina tekstualnih informacija digitalno rođeno te da će kroz nekoliko godina to biti slučaj i za slikovne podatke.

U veljači 2001. godine, analitičar Doug Laney, objavljuje istraživanje [6] u kojem opisuje tri specifične karakteristike velikih skupova podataka, a to su volumen, brzina i raznolikost. Desetljeće kasnije, „3V" koja se odnose na engleske nazive *volume, velocity, variety*, postaju opće prihvaćene tri definirajuće dimenzije velikih skupova podataka.

Danah Boyd i Kate Crawford, 2012. godine objavljuju definiciju pojma „*Big Data"* [7] opisujući ga kao kulturni, tehnološki i znanstveni fenomen koji počiva na međusobnom djelovanju:

- **Tehnologije** - maksimiziranje računalne snage i točnosti algoritama kako bi se skupili, analizirali, spojili i usporedili veliki skupovi podataka
- **Analize** - crtanje na velikim skupovima podataka kako bi se identificirali obrasci za izradu ekonomskih, socijalnih, tehnoloških i legalnih zahtjeva
- **Mitologije** - rašireno mišljenje da veliki skupovi podataka nude viši oblik inteligencije i znanja koji pružaju pronicljivost koja je prije bila nemoguća.

2.2.2. Moderna definicija *VSP-a*

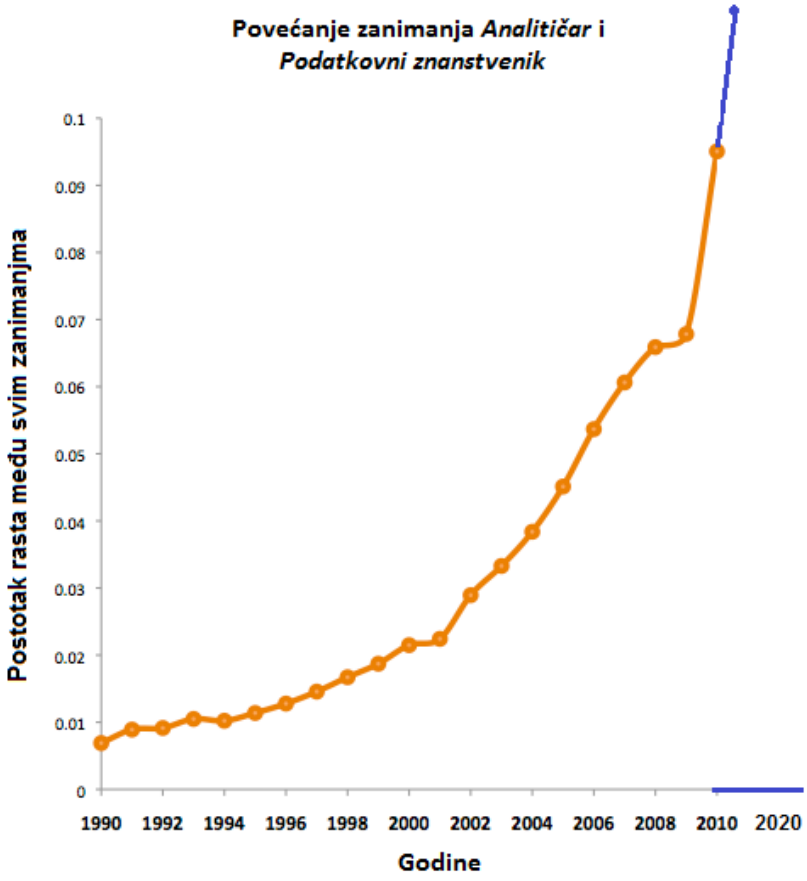
Veliki skupovi podataka je pojam koji opisuje velike količine strukturiranih ili nestrukturiranih podataka s kojima je vrlo teško ili praktično nemoguće raditi korištenjem

standardnih alata ili relacijskih baza podataka [8]. Veliki podaci su sve ono što ne stane u *MS Excel* [9].

Kako bi laici mogli lakše razumjeti kakvi su to podaci koji se nazivaju velikim skupovima podataka, trgovci koji koriste digitalnu tehnologiju za prodaju, podijelili su ih po vlastitom iskustvu u četiri kategorije [10]:

1. **Podaci o kupcu** (*engl. customer data*) – svi kontakt podaci trenutnih kupaca (prošli i sadašnji), podaci o očekivanim kupcima (ljudi koji žele surađivati s vašom firmom, ali još nisu), mail-liste, podaci iz službe za korisnike, itd.
2. **Podaci o kupovini** (*engl. purchase data*) – svi transakcijski podaci kupaca koji su kupili vaš proizvod ili uslugu
3. **Društveni podaci** (*engl. social data*) – bilo koji podatak koji se nalazi na društvenim mrežama, blogovima, forumima ili stranicama koje prikupljaju druga mišljenja
4. **Podaci spajanja** (*engl. connected data*) – svaki spremljeni podatak s bilo kojih spojenih uređaja na internetu – često nazvani **Internet objekti** *(engl. Internet of Things),* a neki od njih su spojeni hladnjak, četkica za zube, *PlayStation*, klima, automobil ili kućna tehnologija kao što je *Nest.*

Razvojem tehnologije raste i količina podataka koja se generira, registrira i sprema u raznim sustavima. Veliki podaci trenutno ulaze u fazu u kojoj eksperimentalna rješenja i ideje počinju postajati pravi produkti u punom smislu te riječi te samim time dobivaju svoju poziciju i primjenu u praksi [11]. Ne tako davno, Shoshana Zuboff je u svojoj knjizi [12] predvidjela da će zaposlenici morati razvijati nove vještine i znanja u skladu s razvojem novih tehnologija koje često podrazumijevaju potpuno novi način razmišljanja. Generalno gledano, razvojem novih tehnologija uvijek su se javljala nova radna mjesta i nova radna zanimanja. Prema predviđanjima Shoshane Zuboff, mi smo ti koji su trenutno u „*Big Data"* eri koja zahtjeva posebna znanja, stručnost i vještine, a zanimanja poput „*Data Scientist, Business Intelligence Consultant, Data Engineer, Big Data Consultant"* i slična, su deficitarna zanimanja budućnosti. Uz trenutnu stopu potražnje spomenutih zanimanja, lako je zaključiti da će se taj trend nastaviti, kao što je prikazano na s*lici 2.2* [13].



Slika 2.2. Grafički prikaz trenda potražnje stručnjaka u području velikih skupova podataka

2.2.3. Glavne dimenzije velikih skupova podataka

Često se spominju dimenzije velikih skupova podataka i vode se rasprave što je dovoljno veliko da bi bilo „veliki podatak“ (je li to relacijska baza podataka od 10 TB ili nešto još veće?). Veličina, odnosno **obujam** podataka samo je jedna od tri glavne karakteristike [14]. Preostale dvije su **raznolikost** i **brzina –** popularno nazvane **3V** zbog engleskih naziva (***volume, variety, velocity***).

Obujam (*engl. volume*) predstavlja veliku brzinu rasta količine novih podataka, a čuvanje postojećih dovodi do toga da se dnevno pohranjuju petabajti (PB) podataka. Kroz nekoliko navedenih primjera može se uočiti njihova veličina, a to su:

- *Twitter* - dnevno generira oko 7 TB podataka, a *Facebook* oko 10 TB
- Američkom magazinu *TIME* 1965. godine, trebala je jedna godina kako bi objavio 50 milijuna riječi. *Twitter* to danas objavi za 8 minuta i 40 sekundi
- Svake minute, isporučeno je 204 milijuna e-mail pošte
- Svake minute, postavljeno je 8 sati video sadržaja na *You-Tube* kanal.

Raznolikost (*engl. variety*) – današnji podaci često dolaze u oblicima koji nisu „uredni“ i strukturirani na način na koji smo navikli. Više nije dovoljno čuvati samo strukturirane podatke, već i slike, podatke s društvenih mreža, logoe, senzorske podatke, itd.

Brzina (*engl. Velocity*) kojom pristižu novi podaci u realnom vremenu je iznimno velika, a veća je od brzine obrade podataka. Doslovno možemo govoriti o „*streaming-u* podataka“. Za primjer se može spomenuti kako se na *Twitter-u* objavi 6000 *tweet*-ova u sekundi.

2.2.4. Dodatne karakteristike velikih podataka

Uz glavne dimenzije, vodeća tvrtka za analitička programska riješenja SAS, dodaje još tri bitne karakteristke [15] za velike skupove podataka, a to su:

- **Varijabilnost (*engl. Variability*)** – tijekom vremena, može se pokazati nedosljednost podataka – što ometa učinkovit proces rukovanja i upravljnja podacima.
- **Složenost (*engl. Complexity*) –** upravljanje podacima može biti vrlo složeno, pogotovo kada velike količine podataka dolaze s više izvora. Podaci moraju biti spojeni i povezani u korelaciju kako bi korisnici shvatili informaciju podatka koji treba biti prenešen.
- **Vjerodostojnost (*engl. Veracity*) –** kvaliteta prikupljenih podataka može biti jako različita (nekonzistentni podaci, dvosmisleni). Točnost analiza ovisi o vjerodostojnosti izvornih podataka.

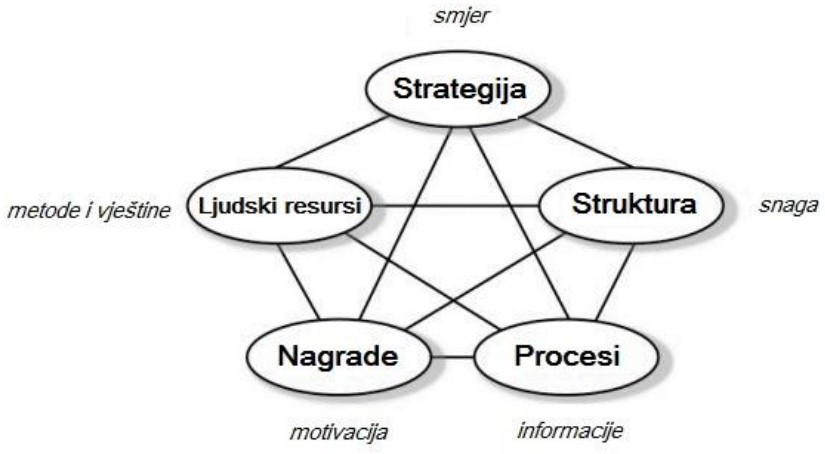
2.2.5. Utjecaj VSP-a na poslovanje

Pod pritiscima dinamičnog i konkurentskog okruženja, organizacije moraju biti dizajnirane tako da svojom efikasnošću ostvaruju vrhunske rezultate, a da istovremeno budu dovoljno fleksibilne i prilagodljive [16]. Bit velikih skupova podataka je transformiranje organizacije u prediktivnu, podacima vođenu organizaciju, koja pomoću podataka u stvarnom vremenu donosi odluke s ciljem poboljšanja svojih poslovnih rezultata. Kako bi organizacija stekla i zadržala konkurentsku prednost, mora kreirati jedinstven pristup prema kupcima, proizvodima i poslovnim procesima. Tu dolazi do izražaja organizacijski dizajn. Organizacijski dizajn predstavlja proces konfiguriranja organizacijske strukture, procesa, sustava nagrađivanja i ljudskih resursa kao ključnih elemenata svake organizacije.

Model organizacijskog dizajna može se definirati kao okvir koji sadrži skup svih komponenti koje se smatraju ključnima za funkcioniranje svake organizacije i prikaz njihovih uzročno-posljedičnih veza. Komponente organizacije koje su izložene utjecaju velikih podataka, u praksi nije jednostavno objasniti. Zbog toga teoretičari u svijetu najčešće primjenjuju Jay R. Galbraith-

ov model zvijezde koji eksplicitno prikazuje ključne komponente organizacije čiji dizajni moraju biti usklađeni, kako međusobno, tako i s okolinom [17].

U osnovi modela zvijezde nalazi se pet ključnih kategorija koje se jasno prepoznaju u svakoj organizaciji. Prema *slici 2.3* to su: strategija (*engl. strategy*), stuktura (*engl. structure*), procesi (*engl. processes*), sustav nagrađivanja (*engl. rewards*) i ljudski resursi (*engl. people*). Strategija određuje smjer, a struktura određuje mjesto odlučivanja. Procesi su povezani s protokom informacija i označavaju sredstva reagiranja prema informacijskim tehnologijama. Sustav nagrađivanja utječe na motiviranost djelatnika da obavljaju i adresiraju organizacijske ciljeve dok se ljudski resursi sastoje od seta ideja i planova koje utječu i definiraju zaposlenikove umne metode i vještine.



Slika 2.3. Prikaz Jay R. Galbraith-ovog Modela zvijezde

Lideri mogu utjecati na sve navedene komponente i oblikovati ih svojim aktivnostima, a one će posljedično djelovati na ponašanje zaposlenika. Kultura kao značajna komponenta organizacije nije obuhvaćena ovim modelom, jer lideri nemaju direktan utjecaj na nju, već je indirektno oblikuju kroz prije spomenutih pet komponenti modela.

Strategija je ključna komponenta svake kompanije i predstavlja ciljeve koje želi postići kao i plan s aktivnostima pomoću kojih će ostvariti te ciljeve.

2.2.6. Utjecaj VSP-a na strategiju

Ukratko rečeno, strategija je željena, ali realna budućnost tvrtke. Željena, jer odražava ciljeve koje menadžment želi ostvariti strategijom, a realna zbog uzimanja u obzir svih faktora iz okruženja i procjena ima li tvrtka dovoljno potencijala za ostvarivanje svojih želja [18]. U tvrtkama postoje tri razine na kojima se donose strategije:

1. Strategija za razinu poduzeća (*engl. corporate-level strategy*) – kojom se određuje budući razvoj tvrtke, odnosno poslovno područje u kojem će obavljati svoje poslove.
2. Strategija za razinu poslovnih jedinica (*engl. business-level strategy*) – koja predstavlja organizacijski podsustav koji ima svoje okruženje i konkurentsko okruženje s kojim se suočava. Svaka poslovna jedinica usvaja vlastitu strategiju koja mora biti u skladu sa strategijom na razini tvrtke.
3. Strategija za razinu funkcionalnih jedinica (*engl. functional-level strategy*) – gdje menadžeri funkcionalnih jedinica donose strategije kako bi realizirali ciljeve koji su im zadani.

Veliki podaci imaju utjecaj na formuliranje sve tri vrste strategija s obzirom da one podrazumijevaju analizu okruženja, identificiranje mogućnosti, vrednovanje i izbor najbolje strategijske alternative. Tvrtke koje uvode tehnologije vezane za velike podatke, formuliraju strategije kojima nastoje izgraditi ključne sposobnosti za brži i kvalitetniji proces odlučivanja kako bi uspjele kreirati vrijednost na osnovu velikog priljeva podataka s kojima su suočene. Prikupljeni podaci provučeni kroz različite programe i obojani različitim bojama, teoretski ništa ne znače ukoliko se na osnovu njih ne donesu nekakve poslovne odluke. Osluškivanje i praćenje zahtjeva, potreba i želja potrošača te prilagodba strategije njihovim potrebama, ključni su uvjeti za uspjeh tvrtke. Primjenom tako orijentiranih tehnologija, tvrtke mogu mjeriti efikasnost svojih marketinških kampanja i inicijativa, preciznije procjenjivati potencijalne rizike, performanse zaposlenika, itd. Prema tome, tvrtka mora imati plan prema kojem će se prikupljati željeni podaci, obrađivati ih i pomoću njih donositi odluke. Eksplozija digitalnih podataka donosi velike izazove, od kojih je jedan od najbitnijih i najtežih – koje podatke prikupljati?

Kada se detaljnije analiziraju poslovne funkcije nabave, proizvodnje, marketinga, prodaje, financija, informatičkih tehnologija, ljudskih resursa, istraživanja i razvoja, nameće se zaključak da svaka od tih grana može imati velike koristi od primjene velikih podataka. Jedna od prvih funkcija koja se mijenja zbog uvođenja VSP-a u poslovanje jest funkcija informatičkih tehnologija. Nova strategija najčešće priprema plan u kojem se utvrđuje što se mora promijeniti kako bi se implementacijom novih tehnologija ostvarili ciljevi tvrtke. Tu uglavnom spadaju odluke o neophodnoj infrastrukturi – sklopovlju i programskim sustavima, vremenu potrebnom za implementaciju i troškovima. Samim time, formuliraju se i strategije koje određuju kako pomoću implementirane tehnologije prikupljati, obrađivati, prikazivati i čuvati podatke. Prva funkcija koja uočava potrebu primjene VSP-a je marketing. Marketing podrazumjeva četiri

koraka: analizu potencijalnih kupaca, privlačenje njihove pozornosti, postizanje interesa kupaca i prihvaćanje postojeće ponude. Sva četiri koraka ovise o marketinškim aktivnostima organizacije. VSP nudi velike potencijale, jer obuhvaća sve podatake s društvenih mreža, blogova, umreženih uređaja koji odražavaju potrebe i želje potrošača, njihove komentare, eventualne primjedbe i procjenjuje njihove potrebe, navike, želje i interese.

2.2.7. Primjena VSP-a u industriji

Kako tvrtke postaju sve više ovisne o količini podataka koje mogu prikupiti i analizirati, tako traže bolje načine za obradu velikih skupova podataka. Analizirajući velike skupove podataka, tvrtke vrlo brzo dobivaju pomoć i bolji uvid u izgradnji jedinstvenog mjesta unutar industrije [19]. Jedna od prvih koja je počela primjenjivati dobrobiti analize velikih podataka jest farmaceutska industrija, odnosno **zdravstvo**. Daleko više medicinskih informacija moguće je prikupljati i analizirati u stvarnom vremenu što liječnicima omogućuje bolju skrb za bolesnika. Koordinacija podataka iz medicinske dokumentacije i usporedba s medicinskim istraživanjima, esencijalno su bitne za bolnice, liječnike i laboratorije. Pomoću međusobno spojenih zdravstvenih uređaja koji su sada i spojeni na internet, znanstvenici povezuju prethodno nestrukturirane skupove podataka, što konstantno dovodi do novih otkrića u liječenju. Tako VSP igraju glavnu ulogu u zdravstvenoj industriji, a uspjeh te misije ovisi o uspješno interpretiranim velikim skupovima podataka i njihovoj primjeni. Veliki skupovi podataka mogu utjecati i na stav **telekomunikacijskih** korisnika tako što pomoću njih pomažu pružateljima usluga dostaviti personalizirano iskustvo zadovoljnog korisnika. Takav način poslovanja dovodi do stjecanja novih pretplatnika, rasta postojećih veza te zadržavanja dragocjenih korisnika. Prema *IBM-ovom* istraživanju, telekomi koji su uveli VSP projekte, doživjeli su pad utrošenog vremena pri obradi i analizi mreže i podataka poziva u iznosu od 92%. Uz bolju analitičku učinkovitost, dolazi do poboljšanja usluga, zadovoljstva kupaca i njihove odanosti. Jednake ciljeve imaju i **financijske ustanove (banke).** Kako bi zadržale vjernost klijenata, potrebne su konstantne prilagodbe njegovim potrebama kao i programi dodatnih pogodnosti. Bankarski stručnjaci stalno osluškuju stanje na tržištu i pomoću analize transakcijskih i drugih podataka mogu predvidjeti ponašanje i potrebe klijenata. Dok upravljaju našim novcem i brinu o zaštiti osobnih financijskih podataka, moraju koristiti posebne sigurnosne sustave. VSP poboljšavaju sigurnosni sustav banaka tako što analizirajući mrežno ponašanje pronalaze sumnjivo ili nenormalno ponašanje. Osim pomoći pri poboljšanju računalne sigurnosti, VSP analize mogu poboljšati sposobnosti financijskih ustanova pri izračunima kreditnih mogućnosti, postavljanju kamatnih stopa i predviđanjima koji su kupci

u opasnosti od primjerice, neplaćanja kreditnih rata. **Turistička industrija** također koristi mogućnosti predviđanja pri korištenju velikih skupova podataka. Prema prošlogodišnjim podacima moguće je zaključiti koje su destinacije bile najposjećenije i zašto, te nekim drugim lokacijama pomoći pri boljoj reklami. Isto tako, analizirajući podatke sa svjetskih tražilica, moguće je predvidjeti poželjne destinacije u budućnosti. Turističke agencije, koristeći mogućnosti VSP-a, smanjuju troškove i povećavaju putničko zadovoljstvo koristeći postojeće podatke o boljim putnim pravcima i vremenskim obrascima, obavijestima o troškovima goriva, ulaznicama za društvene sadržaje ili dostupnosti smještaja. Ti navedeni detalji, omogućuju poboljšanje logistike, sigurnost i zadovoljstvo turista.

2.2.8. Prednosti i nedostaci korištenja VSP-a

Imati puno podataka i informacija u kompaniji ili vlastitim resursima je jedna mogućnost, ali biti u stanju pohraniti ih, analizirati i vizualizirati u stvarnom vremenu (*engl. real-time*) je sasvim drugačija dimenzija posjedovanja podataka [20]. Sve više organizacija žele imati uvid u vlastite procese u stvarnom vremenu kako bi u potpunosti razumjele što se oko njih događa. Postoji mnogo prednosti korištenja VSP analitike u stvarnom vremenu, a neke od njih su:

- Pogreške unutar organizacije su odmah prepoznate. „*Real-time"* uvid pomaže, odnosno omogućuje brzu reakciju tvrtki u ublažavanju učinaka ili uklanjanju problema. To dovodi do osiguranja sustava od kompletnog ispada iz rada ili spašavanja prekida isporuke usluga prema klijentima.
- Nove strategije konkurentske kompanije su odmah vidljive. S VSP analitikom u stvarnom vremenu uvijek možete biti korak ispred konkurencije ili možete biti obaviješteni istog trenutka kada vaša konkurencija promjeni strategiju (npr. snižavanje cijene usluga).
- Usluge se dramatično poboljšavaju što može dovesti do veće stope pretvorbe ili dodatnog prihoda. Kada organizacije prate koje usluge korisnici najčešće koriste, one mogu proaktivno reagirati na moguće dolazeće kvarove. Za primjer možemo prikazati senzore u automobilu. Oni mogu obavijestiti vozača o mogućem budućem kvaru, prije nego se dogodi te tako upoznati vozača s potrebama održavanja njihovog vozila.
- Prijevare mogu biti otkrivene u istom trenutku kada se dogode te se tako mogu poduzeti odgovarajuće mjere ograničavanja štete. Financijski svijet je vrlo privlačan kriminalcima, a s VSP sustavom sigurnosti, pokušaji napada su odmah otkriveni, jer IT odjeli reagiraju odgovarajućim mjerama na vrijeme.

- Ušteda – implementacija VSP analitičkih alata može biti skupa, ali vremenom se uštedi puno više novca. Lideri više ne moraju čekati na izvještaje i popunjavanje baza podataka (korisno za analitiku u stvarnom vremenu). Smanjuje se teret cjelokupne IT podrške oslobađajući resurse koji su se prethodno koristili na zahtjeve za izradu izvještaja.
- Pažljivije motrenje prodaje dovodi do dodatnih prihoda. Analizom se dolazi do zaključaka kako se točno razvija prodaja određenog proizvoda te ako mu je prodaja iznimno dobra, internet trgovac može poduzeti određene preventivne mjere kako bi spriječio nedostatak proizvoda na tržištu i gubitak prihoda.
- Držanje trenda s kupcima – uvidom u konkurentske ponude, promocije ili kretanja kupaca, pružaju se vrijedne informacije o trenutnim i dolazećim trendovima. Analitika pomaže pri bržem donošenju odluka koje usluge bolje odgovaraju trenutnim kupcima.

Iako puno ljudi smatra kako će veliki skupovi podataka imati bitan utjecaj na društvo u različitim segmentima, ipak, postoji skupina onih koji vjeruju suprotno. Kritičari tvrde da VSP djeluju nauštrb svih nas, a jedan od glavnih argumenata je prikupljanje i pohranjivanje osobnih podataka, odnosno narušavanje privatnosti. Korištenjem kreditnih kartica, sustavi prikupljaju sve pojedinosti vezane za kupovinu, od imena osobe koja je koristila karticu, predmeta koji je kupljen, mjesta i vremena kada je predmet kupljen. Imajući pristup takvim informacijama, vladajuće institucije posjeduju svu moć u praćenju i kontroli ljudi. Samim time, postoji i velika mogućnost krivih procjena i (ne)opravdanih sumnji o ilegalnim radnjama. Nezagovornici zasigurno imaju čvrste argumente, a kao što je već spomenuto, glavni nedostaci VSP-a su [21]:

- Sigurnost i privatnost podataka
- Računalni kriminal
- Problemi s kvalitetom i točnošću podataka
- *„Online“* pristup – ukoliko nemate internet pristup, nemate ni pristup podacima
- Trenutna infrastruktura nije dovoljno dobra, potrebna su računala s većom snagom i memorijskim pristupom
- Potrebne su dodatne investicije za razne programske alate i sustave
- Neadekvatno osoblje, odnosno manjak vještina i znanja kod zaposlenika što dovodi do novih troškova namjenjenih za obrazovanje i trening zaposlenog kadra

2.2.9. Tehnologije VSP

Što su veliki skupovi podataka u tehnološkom smislu i kako se implementiraju, česta su pitanja u poslovnom svijetu. Implementirati VSP ne znači samo postaviti nekakvu bazu i u nju spremati podatke. Za većinu velikih tvrtki to je zapravo *Hadoop* projekt. *Hadoop* je danas implementiran u proizvodima *Cloudera-e, Hortonworks-a, IBM-a, MapR-a, Microsofta* i drugih, a što je *Hadoop* i kako se koristi, bit će objašnjeno u sljedećem poglavlju. Na *slici 2.4* prikazane su razne infrastrukture i usluge koje zajedno čine okolinu velikih skupova podataka [22].

Slika 2.4. VSP tehnologije i alati

3. PROGRAMSKO OKRUŽENJE *HADOOP*

Apache Hadoop je programsko okruženje otvorenog koda (*engl. open-source framework*) koje, koristeći jednostavne programske modele, služi za raspodjeljenu pohranu i obradu velikih skupova podataka na računalnim nakupinama (*engl. clusters*) [23]. Nastao je 2005. godine, a kreirali su ga Doug Cutting i Mike Cafarella u programskom jeziku *Java*. Ime je dobio po žutom slonu, plišanoj igrački Cuttingovog sina.

3.1. Razlozi izbora *Hadoop-a*

Programsko okruženje otvorenog koda, *Hadoop,* osigurava **pohranu velike količine podataka** (bilo koje vrste) uz iznimno **snažnu procesorsku obradu** virtualno neograničenog broja zadataka ili poslova [24].

Program otvorenog koda (*engl. open-source software*) je program kojeg kreiraju i održavaju preko mreže programeri iz cijelog svijeta. Pojam „otvoreni kod“ odnosi se na nešto što se može modificirati i što je zajedničko svima, jer je njegov dizajn javno dostupan [25]. Iako je nastao u kontekstu razvoja računalnih programskih sustava, danas taj termin označava skup vrijednosti koje prihvaćaju i slave otvorenu razmjenu, zajedničko sudjelovanje, brzo prototipiranje, transparentnost i razvoj zajednice i upravo je to jedan od glavnih razloga zašto je *Hadoop* široko prihvaćen. **Programsko okruženje** (*engl. framework*) je pojam koji označava skup alata koji se nalaze na jednom mjestu, a potrebni su za razvoj i pokretanje programskih aplikacija kao što su programi, veze i sl. *Hadoop* sadrži brojne alate za različite potrebe što ga čini idealnim partnerom u radu s Velikim skupovima podataka. Zbog toga je **pohrana velike količine podataka** od velike važnosti. Hadoop okruženje ima originalan pristup pri kojem razbija velike podatke u blokove te ih multiplicira i sprema na poslužitelje. Za istovremenu obradu tako velikih podatka *Hadoop* koristi **snagu procesora** više međusobno povezanih „jednostavnih računala“ koji su financijski dostupni svima.

3.1.1. Prednosti korištenja *Hadoop-a*

Jedan od glavnih razloga zašto tvrtke uvode *Hadoop* u poslovanje je njegova sposobnost pohrane i obrade velike količine podataka bilo koje vrste velikom brzinom. Uz konstantno povećanje obujma i izvora podataka (raznolikost) s društvenih mreža i Internet objekata, brzina obrade je od ključnih osobina koji se uzimaju u obzir. Uz brzinu, ostale pogodnosti su:

- Računalna snaga – *Hadoop-ov* raspodijeljeni računalni model brzo obrađuje podatke. Što se veći broj računalnih čvorova koristi, veća je računalna snaga.
- Fleksibilnost – za razliku od tradicionalnih relacijskih baza podataka, podaci se ne moraju predobraditi prije spremanja. Korisnik može pohraniti koliko želi podataka i poslije odlučiti što želi s njima. To mogu biti tekstualni podaci, slike, video zapisi i razni drugi.
- Toleriranje kvarova – obrada podataka je zaštićena od potencijalnog kvara sklopovlja. U slučaju kvara jednog čvora, poslovi se automatski preusmjeravaju na druge čvorove kako bi se osiguralo raspodjeljeno računarstvo od kvara. Uz to, u slučaju kvara/ispada, sustav automatski sprema kopije svih podataka.
- Niska cijena – okruženje otvorenog koda je besplatno i koristi poslužitelje za pohranu velikih skupova podataka.
- Skalabilnost – nadogradnja sustava se provodi jednostavnim dodavanjem više čvorova u sustav gdje nisu potrebne velike administrativne promjene.

3.2. Arhitektura *Hadoop-a*

Hadoop je dizajniran kako bi pružio usluge prema pojedinačnim poslužiteljima (do tisuće uređaja), pri tome osiguravajući lokalne proračune i pohranu [26]. Kako bi pružio isporuku visoke raspoloživosti, *Hadoop* se ne oslanja na sklopovlje, već na same biblioteke koje su dizajnirane za otkrivanje i otklanjanje kvarova na aplikacijskom sloju i isporuku visoko-raspoloživih usluga na vrhu nakupine računala. Jezgra *Hadoop*-a sastoji se od dijela za pohranu – ***HDFS*** (*engl. Hadoop Distributed File System*) i dijela za obradu – ***MapReduce***. *Hadoop* dijeli datoteke u velike blokove i distribuira ih među čvorovima unutar nakupine računala. Za obradu podataka *Hadoop MapReduce* prenosi zapakirane kodove čvorova kako bi se paralelno obradili prema principu da se svaki čvor mora obraditi. Takav pristup daje prednost u odnosu na lokalnost podataka – čvorovi manipuliraju podacima koje imaju – što omogućuje bržu obradu podataka i veću učinkovitost u odnosu na konvencionalne super-računalne arhitekture koje se oslanjaju na paralelni datotečni sustav gdje su podaci i proračuni povezani mrežom velike brzine (*engl. high-speed network*).

Hadoop se sastoji od 4 glavne komponente. To su:

- *Hadoop Common* paket – sadrži biblioteke i uslužne programe za druge module

- *Hadoop* raspodijeljeni datotečni sustav (*Hadoop Distributed File System - HDFS*) – sustav koji pohranjuje podatke na strojevima za pričuvu
- *Hadoop MapReduce* – programski model za obradu velikih skupova podataka
- i *Hadoop YARN* – platforma odgovorna za upravljanje računalnih resursa u nakupinama računala koja ih koristi kod raspoređivanja korisničkih aplikacija.

3.2.1. *Hadoop Common* paket

Hadoop Common paket sadrži potrebne *Java* arhive odnosno *JAR* datoteke i skripte koje služe za pokretanje *Hadoop*-a [27]. Ovaj paket sadrži izvorni kod i dokumentaciju. Isto tako, u tom paketu se nalaze i svi potrebni elementi za komunikaciju *Hadoop*-a s ostalim alatima. Struktura paketa mijenja se s obzirom na predstavljanje novih verzija. Za korištenje *Hadoop*-a nije potrebno poznavati *Common* paket osim ako se netko ne želi baviti razvojem samog *Hadoop*-a.

3.2.2. *Hadoop* raspodijeljeni datotečni sustav - *HDFS*

HDFS je raspodijeljeni datotečni sustav i *Hadoop-ov* sastavni dio. To je zapravo sustav u kojem je definirano kako se podaci pohranjuju, kopiraju i čitaju [28]. Zaslužan je za mogućnost lakog pohranjivanja velike količine podataka. Dizajniran je tako da može raditi na bilo kojoj sklopovskoj infrastrukturi iako se njegova prava moć vidi na poslužiteljima. Sustav je jako otporan na greške i nije sklopovski zahtjevan. Kako je namjenjen za poslužitelje, točnije za stotine ili tisuće poslužitelja, koji imaju različite komponente i za koje postoji vjerojatnost kvara, glavni cilj *HDFS* arhitekture je brzo otkrivanje pogrešaka te automatsko otklanjanje istih. Namjenjen je za manipulaciju velikom količinom podataka (najmanje u gigabajtima-GB) i ima jednostavan model pristupa datotekama – „upiši jednom – učitaj više puta“. *HDFS* nije namjenjen za interakciju s korisnikom, nego neometanoj obradi podataka od strane drugih alata. Gotovo je nemoguće pristupiti podacima i pročitati ih u razumljivom formatu koji su uneseni prilikom direktnog pristupa stroju na kojem je *HDFS* instaliran.

HDFS ima *master/slave* arhitekturu. Konkretno to znači da se *Hadoop* nakupina računala sastoji od jednog *NameNode*-a i više *DataNode*-ova. Obično je jedan poslužitelj *master* i na njemu se instalira *NameNode*, a na ostalim poslužiteljima *DataNode*-ovi. *Master* poslužitelj na kojem je instaliran *NameNode*, kontrolira pristup datotekama i upravlja prostorom za dodjeljivanje imena (*engl. the file system namespace*) koji podržava tradicionalnu hijerarhiju. Korisnik može kreirati direktorij i u njemu pohraniti datoteku, može mjenjati ime direktorija, brisati ga ili premještati.

Isto to vrijedi i za datoteke. Korisnik također upravlja datotekama na *master* poslužitelju te ima pristup direktorijima i datotekama. Nadalje, datoteke se dijele u blokove i skladište na ostale poslužitelje na kojima je instaliran *DataNode*. *DataNode* služi za pohranu blokova, dozvoljava kreiranje blokova, brisanje i **repliciranje**. Jednostavno rečeno, *NameNode* čuva meta podatke blokova i pomaže krajnjim korisnicima vidjeti datoteku, a ne čuva blokove koji korisniku nisu čitljivi, dok *DataNode* čuva podatke. To se može protumačiti i kao da se u *NameNode*-u čuvaju adrese blokova, ime i broj kopija.

HDFS ima mogućnost upisivanja blokova više puta što smanjuje rizik od gubitka podataka. Najčešće sadrži tri kopije podataka što znači da će se svaki blok kopirati tri puta. To se ne mora dogoditi na samo jednom poslužitelju, nego na svim poslužiteljima unutar grozda na kojima je instaliran *DataNode* (npr. Za 1 TB podataka bit će potrebno 3 TB prostora).

Komunikacija između poslužitelja u grozdu odvija se pomoću *TCP/IP* protokola pa samim time i *NameNode* komunicira s *DataNode-ovima* tako da oni periodički šalju impulse tzv. (*engl Heartbeat*) *NameNode-u*. S obzirom na repliciranje blokova, korisnici su sigurni u slučaju kvara jednog ili čak dva podređena poslužitelja, jer će im uvijek ostati još jedna kopija podataka. Međutim, ako je *master* poslužitelj u kvaru, a korisnik nema meta podatke, sav rad je uzaludan. Iako se i za to traže riješenja (mogućnost kopiranja *master* poslužitelja – kreiranje sekundarnog *NameNode-a*), za sada je potencijalna mogućnost kvara *master* poslužitelja jedina mana *HDFS-a*.

3.2.3. *Hadoop MapReduce*

MapReduce je programski model za obradu velikih količina podataka čiji se algoritam izvršava paralelno i raspodjeljeno (primjerice, kod grozda računala od tri poslužitelja, algoritam će se izvršavati paralelno, odnosno u isto vrijeme na sva tri poslužitelja) [29]. Ovaj model se može promatrati i kao dvije odvojene cjeline, dio *Map* i dio *Reduce*.

Map – služi za jednostavno ili složeno sortiranje i filtriranje podataka (npr. sortiranje studenata po prezimenu).

Reduce – kombinira podatke koje je *Map* obradio (npr. zbrajanje koliko se puta jedna riječ ponovila u zadanim rečenicama).

HDFS i MapReduce – prava moć *MapReduce-a* je u kombinaciji s *HDFS-om*. Svi podaci se pohranjuju kao blokovi na *DateNode-ovima*, a na *NameNode-ovima* se čuvaju meta podaci o tim blokovima. Kada su podaci podjeljeni u blokove, lakše ih je obraditi, stoga je skladištenje podataka u blokove znatno olakšalo *Map* funkciji da ih grupira. Budući da *HDFS* ima više

podatkovnih čvorova (*DataNode*) na kojima se vrši podjela i pohrana podataka na blokove, moguće je iskoristiti računalnu snagu svakog od tih čvorova te provesti zadatke na njima. Dakle, svaki čvor može provesti *Map* ili *Reduce* zadatke, a s obzirom da svaki podatkovni čvor sadrži više podataka, moguće je očekivati izvršavanje zadataka u isto vrijeme za različite podatkovne blokove. Osim njih, bitnu ulogu imaju:

- *JobTracker* – komunicira s *NameNode-om* kako bi dodijelio *MapReduce* zadatke određenom čvoru unutar nakupine računala
- *TaskTracker* – pokreće i prati odvijanje *MapReduce* zadataka u nakupini računala; kontaktira *JobTracker-a* u vezi dodijeljenih zadataka i ako se određeni zadatak ne izvrši, njegov status šalje *JobTracker-u* koji taj isti zadatak dodjeljuje nekom novom čvoru unutar nakupine računala.

3.2.4. *Hadoop YARN*

U prvoj generaciji *Hadoop-a, YARN* (*engl. Yet Another Resource Negotiator*) komponenta nije postojala, već je njen posao bio sastavni dio *MapReduce-a*. *YARN* je uveden kao nova komponenta u drugoj generaciji *Hadoop-a* čiji je cilj bio da se dotadašnji *MapReduce* odvoji u dva dijela kako bi se olakšalo korištenje čitave platforme [30]. Glavna funkcija *YARN-a* je upravljanje resursima u nakipini računala. Može se reći da se sastoji od dvije komponente: *Scheduler* i *ApplicationsManager* koje zajedno čine *Resource Manager*. Izdvajanjem ovog procesa u novu komponentu dovelo je do toga da *MapReduce* služi samo za obradu podataka. Još jedna mogućnost koja se javila s *Hadoop-om* druge generacije, odnosno s *YARN-om*, je to da se sada može pokrenuti veći broj aplikacija koje su pisane za *Hadoop*. To se posebno odrazilo na poslovni svijet, jer se s mogućnošću paralelnog obavljanja više stvari u isto vrijeme stvorila konkurencija na tržištu.

Scheduler je komponenta koja brine o alokaciji resursa aplikacijama koje se izvršavaju. Bitno je napomenuti da se vodi računa samo o resursima, odnosno ne brine se o tome kakav je status aplikacije koja se izvršava, tj. ne prati se rad aplikacije. Kako se brine samo o alokaciji resursa, ne vodi se računa o tome da li je došlo do greške ili je kod loš, što znači da će resursi biti dodjeljeni nekoj aplikaciji dok god se njen rad ne prekine od samog korisnika ili neke druge komponente.

ApplicationsManager upravlja aplikacijama pisanim za *Hadoop*. Njegov zadatak je pregovarati i prihvatiti posao. Pregovarati znači, ispitivati resurse i donositi zaključke što prvo treba izvršiti. Također, zadužen je za resetiranje posla, odnosno aplikacije ukoliko dođe do neke pogreške.

3.3. Ekosustav *Hadoop*

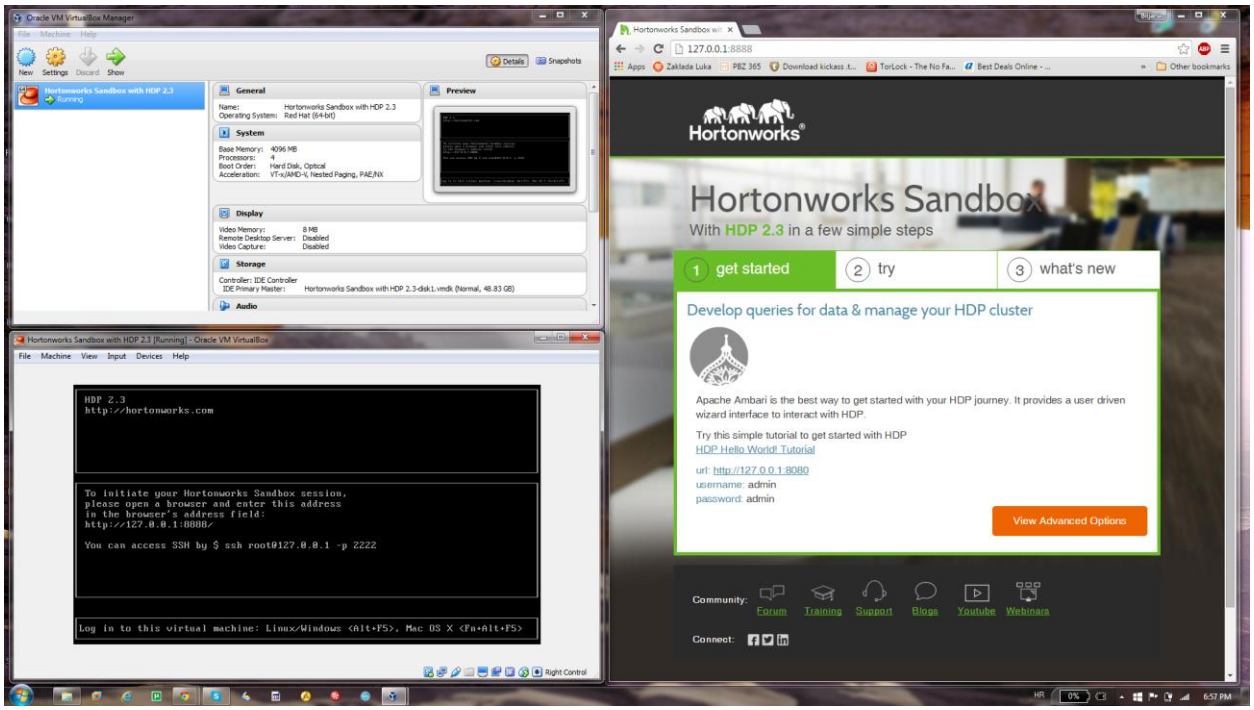
Hadoop ekosustav je skup alata (projekata) koji zajedno rade na *Hadoop* platformi. Među tim alatima su, uz *HDFS* i *MapReduce*, različiti *Apache* licencirani projekti. Na tržištu postoji puno alata koji su u relaciji s *Hadoop-om*, a tvrtke poput *Facebook-a* i *Microsoft*-a razvijaju svoje vlastite koji su također dostupni za instalaciju. Danas postoje i kompanije koje pružaju besplatne usluge, odnosno *Hadoop* platforme s već izgrađenim ekosustavom što olakšava izbor potrebnih projekata za rad i proces instalacije na osobnom računalu. Među najpoznatijima su svakako *Cloudera, Apache, MapR Technologies, IBM* i druge, ali najbolju platformu za početnike nudi besplatna *Hortonworks podatkovna platforma* (*engl. Hortonworks Data Platform - HDP*. Neki od osnovnih alata (projekata) koji su dostupni na različitim platformama su:

- Raspodijeljeni datotečni sustavi *(HDFS)*
- Raspodijeljeno programiranje *(MapReduce, Apache Pig, Apache Tez)*
- *NoSQL* baze podataka *(Apache HBase, Apache Accumulo)*
- *SQL* baze podataka *(Apache Hive, Apache HCatalog)*
- Unošenje podataka *(Apache Flume, Apache Sqoop, Apache Storm)*
- Programiranje usluga *(Apache Zookeeper)*
- Upravljanje podacima *(Apache Oozie, Apache Falcon)*
- Strojno učenje *(Apache Mahout)*
- Sigurnost *(Apache Knox)*
- Razvoj sustava *(engl. System Deployment (Apache Ambari, HUE))*

3.3.1. *Podatkovna platforma Hortonworks – HDP*

Podatkovna platforma Hortonworks je poslovno rješenje tvrtke *Hortonworks* koja je nastala 2011. godine u SAD-u. *HDP* je poduzetnički orijentirana platforma za upravljanje podacima koja omogućuje centraliziranu arhitekturu za pokretanje neizravnih, interaktivnih aplikacija u stvarnom vremenu paralelno s raspodijeljenim skupovima podataka. Izgrađena je na *Apache Hadoop* projektu i podržava sveobuhvatan skup alata koji rješavaju temeljne zahtjeve sigurnosti, poslovanja i upravljanja podacima [31]. Kao što je već prije spomenuto, *HDP* je besplatna platforma, primjenjiva na *Windows* ili *Mac* operacijskim sustavima te zahtjeva minimalno poznavanje programiranja što ju čini savršenim alatom za početnike. Za što lakši početak učenja i programiraja s *Hadoop-om, Hortonworks podatkovna platforma* nudi besplatno preuzimanje i instalaciju *Hortonworks Sandbox* sustava na njihovim službenim stranicama [32]. Jedini preduvjet je da korisnik već ima instaliran *VirtualBox, VMware* ili neki drugi virtualni stroj. Za

sve one koji se po prvi puta susreću s instalacijom programa na virtualnom stroju, *HDP* je također priredio kratak „korak po korak“ vodič za instalaciju *Sandbox*-a [33]. Nakon uspješne instalacije i pokretanja, korisnik se susreće s terminalom koji pruža mogućnost prijave i programiranja u terminalu ili prebacivanja na grafičko korisničko sučelje preko internet pretraživača *(slika 3.1).*

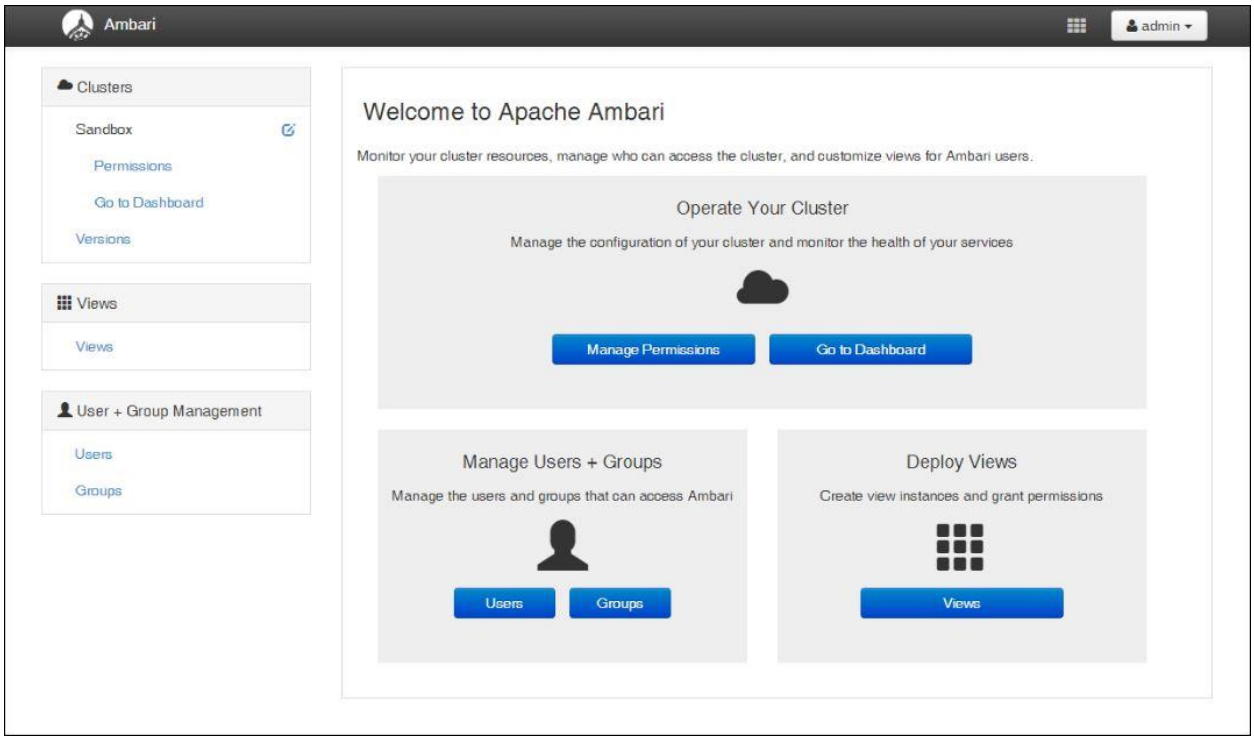


Slika 3.1. *Hortonworks Sandbox* sustav (terminal i grafičko sučelje)

Prelaskom na internet pretraživač, korisnik se susreće s intuitivnim grafičkim sučeljem koje omogućuje lako pretraživanje, komunikaciju i pristup najnovijim informacijama vezanih za *Hortonworks Sandbox*. Jedna od važnih početnih mogućnosti je pokretanje praktičnog vodiča za početnike te prijava u *Ambari* sustav.

3.3.2. Sustav *Apache Ambari*

Apache Ambari projekt, usmjeren je na stvaranje što jednostavnijeg *Hadoop* upravljanja preko razvoja programa za **instalaciju, upravljanje i praćenje** *Hadoop* nakupine računala. *Ambari* pruža intuitivno, jednostavno za korištenje internet sučelje [34]. Nakon prijave u sustav s korisničkim imenom *admin* i lozinkom *admin,* pojavljuje se početna stranica *Ambari* sustava, što se može vidjeti na *slici 3.2.*



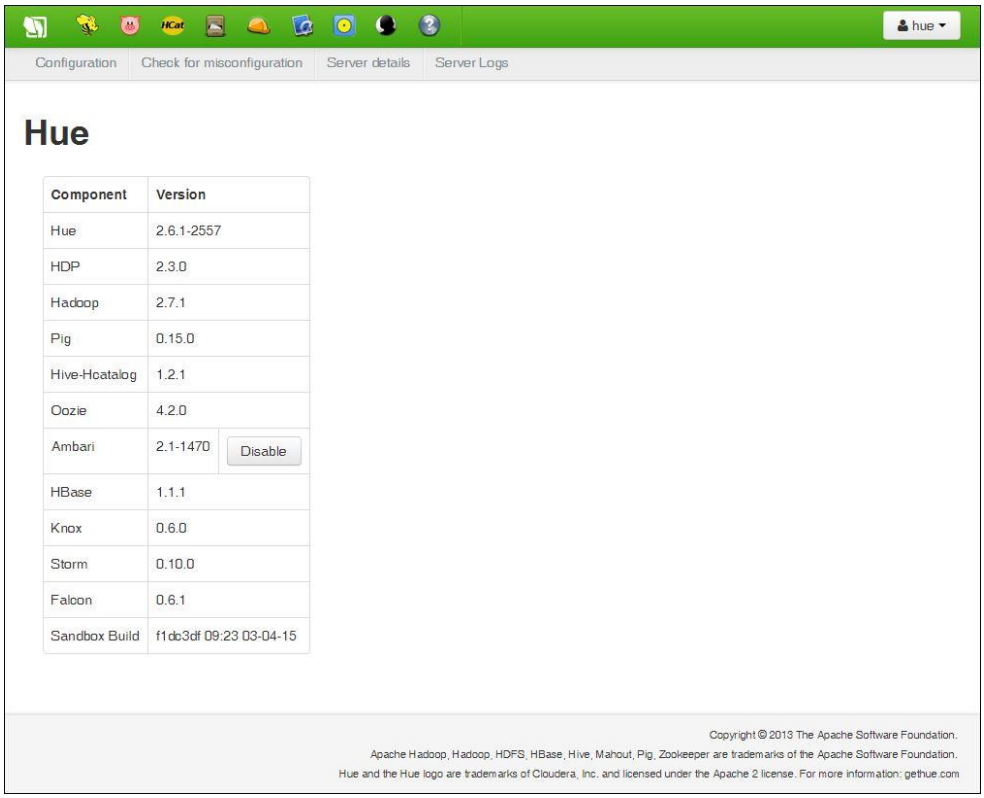
Slika 3.2. Početna stranica *Apache Ambari* korisničkog sučelja

Instalaciju *Hadoop-a* korištenjem *Ambari sustava* moguće je pokrenuti na operacijskom sustavu *Linux*. Na početku je dovoljno instalirati poslužitelj *Ambari* na samo jednom stroju koji čak ni ne mora biti u grozdu, ali je neophodno pokupiti ključeve za autentifikaciju s ostalih strojeva kako bi poslužitelj mogao poslati svoje „agente“ na svaki izabrani stroj. Za instalaciju na *Windows OS-u* nije potrebno koristiti *Ambari*, ali se mogu koristiti sve ostale njegove funkcionalnosti.
Upravljanje *Hadoop* nakupinom računala, pokretanje, zaustavljanje i rekonfiguracija *Hadoop* alata odvija se u samo nekoliko klikova mišem. *Ambari* pruža mogućnost odabira bilo kojeg stroja iz nakupine te pokretanje neke od gore navedenih akcija. S inačicom 1.5.1. dodana je i nova mogućnost, resetiranje nekog alata što se ranije radilo na način „zaustavi pa pokreni“. Još jedna zanimljiva mogućnost je da se neki projekt može implementirati u „*Maintenance Mode*“.
Ambari pruža odlično sučelje za **praćenje** bitnih funkcionalnosti grozda računala. S lakoćom se može vidjeti koliko prostora je preostalo na grozdu, kakvo je stanje *NameNode-ova* i slično.

3.3.3. Korisničko sučelje *Hue*

Hue je internet korisničko sučelje otvorenog koda za *Hadoop* i njegov ekosustav prikazan na *slici 3.3.* Napisan je u *Pythonu* i podržava najčešće alate iz ekosustava. Ako se korisnik želi baviti samo analizom podataka, tada je *Hue* odličan izbor, jer nema korištenja terminala i komandne linije. U slučaju da je korisnik administrator *Hadoop* grozda ili programer, tada *Hue*

nije dovoljan, jer ne podržava sve alate i neke se stvari ipak trebaju odraditi preko *Linux* terminala.



Slika 3.3. *Hue* grafičko korisničko sučelje

Alati koje korisnik može koristiti preko *Hue* grafičkog sučelja su:

- Podatkovni preglednik – (*engl.file browser)*
- *Apache Hiv*e
- *Apache* Pig
- *Apache HCatalog*
- *Apache Oozie* te drugi podatkovni preglednici.

3.3.4. *Apache Hive*

Apache Hive je infrastruktura koja se koristi za skladištenje i obradu velike količine podataka na *Hadoop-u*. *Hive* pruža nešto što se naziva *HiveQL*. Može se reći da predstavlja standard za *SQL* upite nad velikim količinama podatka i još uvijek je u razvoju. Lako se može integrirati s postojećim alatima korištenjem *JDBC* ili *ODBC* sučelja, tj. moguće ga je povezati s *Microsoft Excel-om* i sličnima. Karakterizira ga organizacija i pohrana velikih skupova podataka iz različitih izvora te pružanje korisnicima mogućnost pretraživanja, strukturiranja i analize

podataka za poslovnu inteligenciju (*engl. business intelligence - BI*) [35]. Način rada *Hive-a* je jako sličan relacijskom modelu. Tablice su slične tablicama relacijskog modela, a podaci su organizirani od većih prema manjim jedinicama. Podacima se pristupa upitima koji su slični *SQL-u*. Za razliku od poznatih *SQL* baza podataka, *Hive* ne podržava brisanje i ažuriranje podataka. Razlog tomu je što se podaci u *HDFS-u* mogu samo upisivati, ali ne i mjenjati, a s obzirom da *Hive* radi na *Hadoop-u*, kojem je *HDFS* sastavni dio, logično je zaključiti zašto je to tako.

3.3.5. *Apache HCatalog*

Apache HCatalog je alat koji omogućava lakše upravljanje skladištenjem podataka i tablica na *Hadoop-u* te pruža korisnicima lakše upisivanje i čitanje podataka. U praktičnom smislu, *HCatalog* predstavlja sloj na *Hadoop-u* koji omogućava prikazivanje podataka s *HDFS-a* u obliku tablica. Samim time korisnici nemaju potrebu brinuti se gdje i u kojem formatu su podaci sačuvani. *HCatalog* podržava čitanje i pisanje datoteka u formatima za koje je moguće napisati *Hive SerDe* (*Serializer-Deserializer*), a to su *RCFile* format, *CSV, JSON*, tekstualni, slijedni/sekvencijalni i *ORC* format [36].

3.3.6. *Apache Pig*

Apache Pig je platforma koja *Hadoop* korisnicima omogućuje pisanje složenih *MapReduce* transformacija pomoću jednostavnog skriptnog jezika *Pig Latin. Pig* prevodi *Pig Latin* skriptu u *MapReduc*e program koji se izvršava na *YARN-u* kako bi imao pristup podacima na *HDFS-u* [37]. *Pig* je napravljen tako da izdvoji *MapReduce* kod napisan u programskom jeziku *Java.* Za razliku od *SQL-a* koji je deklarativni programski jezik, *Pig* je slijedni jezik što znači da način na koji se program piše, definira kako će se podaci transformirati. Skripte napisane u ovom jeziku mogu biti grafovi što znači da je moguće pisati složene transformacije s više ulaza i izlaza. *Pig* može raditi u dva stanja, lokalnom i *MapReduce* stanju. Ovaj alat namjenjen je prvenstveno za *ETL* (*engl. Extract-Transform-Load*) poslove za rad nad sirovim podacima i iterativno procesiranje podataka.

U sljedećem poglavlju ukratko će biti prikazana primjena *Hadoop* ekosustava kroz prethodno opisane alate. Nakon instalacije potrebnog programskog okruženja, koristeći objašnjenja i upute koji su dani u sljedećem poglavlju, korisnici će dobiti kratak uvid u velike skupove podataka i kako izvršiti određene manipulacije na *Hortonworks* podatkovnoj platformi koristeći *Pig, Hive, HCatalog i MS Excel.*

4. PRIMJENA EKOSUSTAVA *HADOOP*

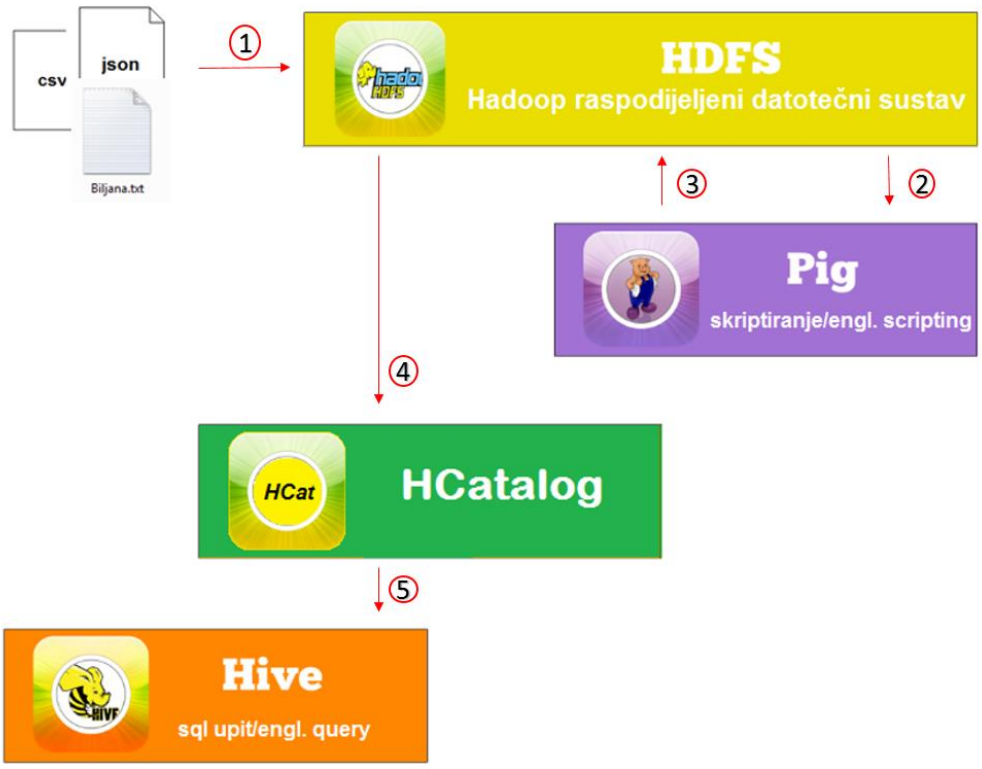
Cilj ovog poglavlja je, kroz dva testna primjera, prikazati prethodno opisane alate *Hortonworks* podatkovne platforme (*engl. Hortonworks Data Platform - HDP*) koja služi za pohranu i obradu podataka na *Hadoop* programskom okruženju te njihovu svrhu.

4.1. Program prebrojavanja riječi u tekstu *Pig*

Svaki programer početnik se barem jednom u životu susreo s primjerom *„Hello World“* pri učenju nekog novog programskog jezika. Primjer koji slijedi nije tipičnog *„Hello World“* karaktera, ali se svakako, zbog njegove jednostavnosti, vrlo često koristi pri upoznavanju korisnika sa stilom programiranja u *Pig Latin* jeziku.

4.1.1. Potrebni alati za ostvarenje programa *Pig*

U ovom primjeru koristit će se proizvoljna tekstualna datoteka (može se koristiti i *csv, json* datoteka, baza podataka ili nešto slično), podatkovni preglednik, *HDFS*, *Apache Pig*, *Apache HCatalog* i *Apache Hive* na *Hue* korisničkom sučelju. *Slika 4.1* prikazuje korake korištenja alata za uspješnu izradu prvog primjera.



Slika 4.1. Slijed korištenja alata za izradu prvog primjera

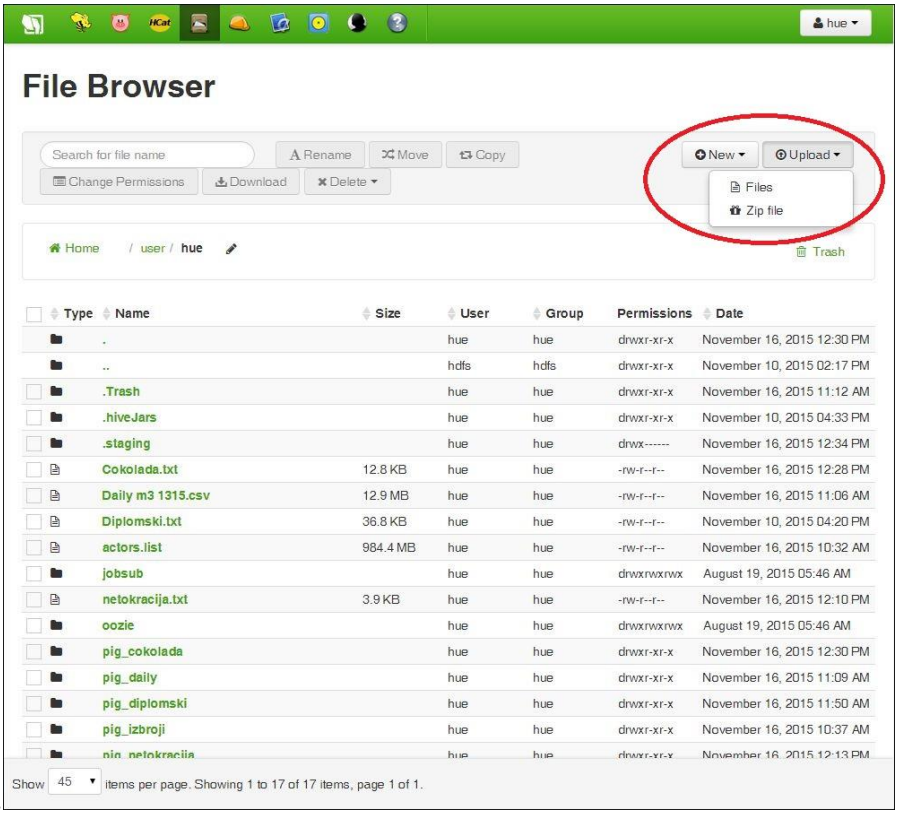
4.1.2. Ostvarenje programa za prebrojavanje riječi

Prvo što je potrebno napraviti jest kreirati (ili preuzeti s Interneta) i spremiti nekakav podatkovni dokument u *Sandbox-ov HDFS* sustav. Zbog jednostavnosti, poželjan bi bio tekstualni dokument (*.txt* dokument), ali mogu se koristiti baze podataka i slično. U ovom primjeru koristit će se sljedeći tekst:

- *Biljana Krpan je studentica. Biljana trenutno pise diplomski rad. Diplomski rad ima temu "Analiza velikih skupova podataka u oblaku racunala". Jako je zanimljivo pisati diplomski rad. Dok Biljana pise diplomski, voli jesti cokoladu. Svi vole cokoladu. Mogu li pronaci cokoladu u oblaku?*

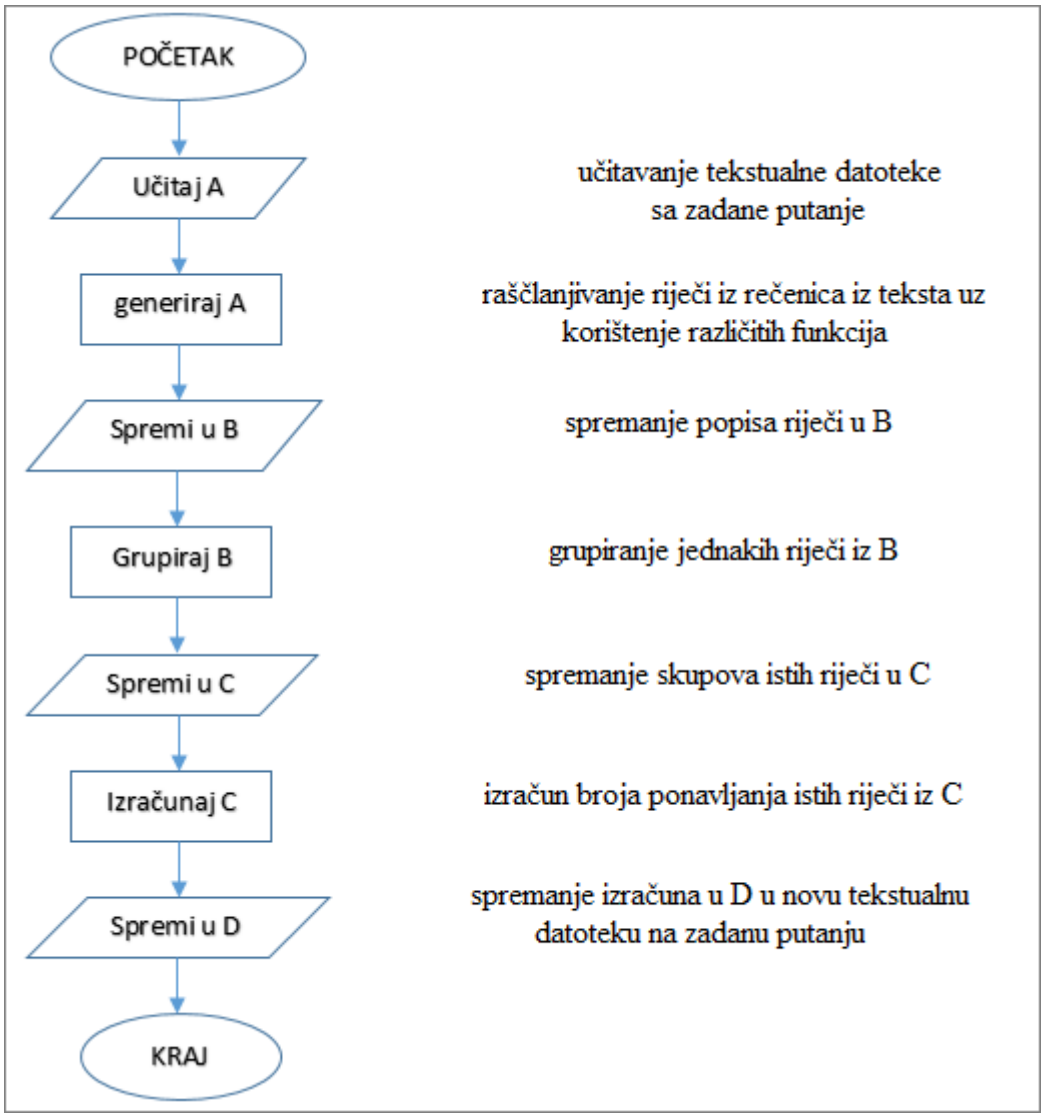
Može se primjetiti da su u tekstu izostavljena hrvatska slova (č,ć, ž, š), jer ih sustav ne podržava. Isto tako, rečenice su namjerno osmišljene u obliku da se ponavljaju iste riječi u istom padežu zbog lakšeg dobivanja rezultata brojanja riječi, jer bi ukupni rezultat pojavljivanja za obje riječi „čokolada“ i „čokolade“ bio jedan, a ne dva. Kako bi se kroz primjer lakše došlo do rezultata, preporuča se koristiti tekstove na engleskom jeziku.

Najlakši način pohrane podataka u *Hadoop HDFS* sustav odvija se preko ***Hue*** grafičkog sučelja kroz ***File Browser/Upload/Files*** putanju što se može vidjeti na *slici 4.2.*



Slika 4.2. Spremanje datoteke u *HDFS* preko *Hue* grafičkog sučelja

Nakon uspješnog kreiranja i spremanja tekstualne datoteke, slijedi kratka *Pig* skripta koja izračunava broj ponavljanja istih riječi u zadanom tekstu. Za ispravan program/skriptu potreban je ispravan algoritam koji se kasnije može jednostavno implementirati u odabrani programski jezik. Razrađeni algoritam i dijagram toka skripte prikazani su na s*lici 4.3.*

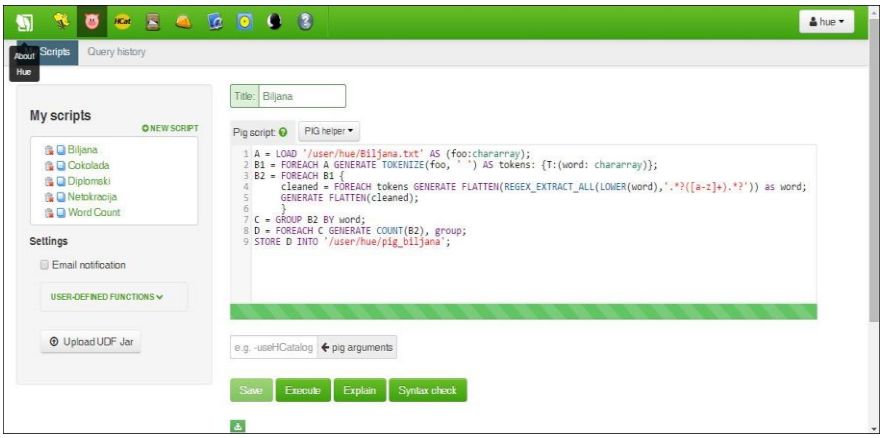


Slika 4.3. Dijagram toka programa za izračunavanje broja riječi

Kodni prikaz algoritma unutar *Pig sustava* može se vidjeti na *slici 4.4,* a za razumijevanje svih korištenih naredbi i funkcija, slijedi detaljan opis po svakoj liniji skripte:

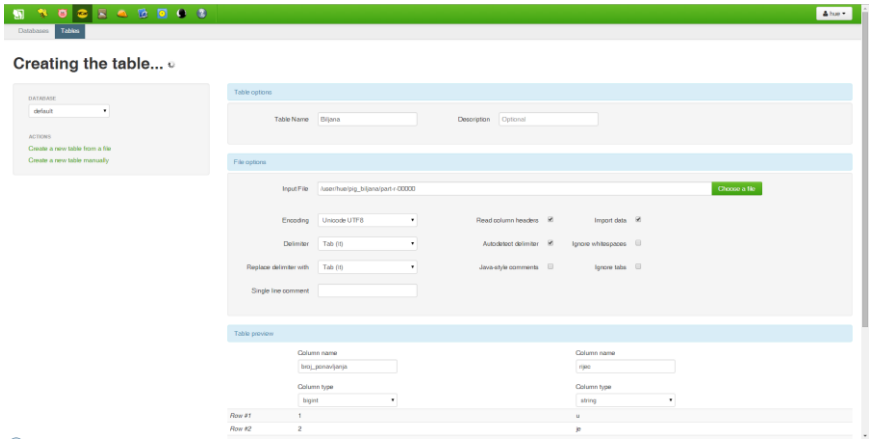
- Linija A – učitava se tekstualna datoteka *Biljana.txt* s mjesta gdje je pohranjena (iz *HDFS-a*)
- Funkcijom *TOKENIZE* rečenice iz teksta se raščlanjuju na riječi (linija B1)

- Raščlanjene riječi se odvajaju u stupac jedna ispod druge (funkcija *FLATTEN*), dok se istovremeno pomoću *LOWER(word)* sve riječi prebacuju u mala početna slova, a s *REGEX_EXTRACT_ALL* odbacuju interpunkcijski znakovi uz zadnje riječi u rečenici (linija B2)
- C linija grupira iste riječi unutar tablice riječi
- D linija prebrojava koliko su se puta iste riječi ponovile, zatim te iste ponovno grupira
- Naredbom *STORE*, rezultat se sprema u direktorij *pig_biljana*



Slika 4.4. *Pig* skripta za prebrojavanje riječi

Kako bi *SQL* programeri mogli što lakše ispitivati podatke na njima razumljiv način, rezultate iz prethodne datoteke potrebno je prebaciti u tablicu. To se vrlo lako može napraviti u *HCatalog-u*. *Slika 4.5* prikazuje kreiranje nove tablice pod imenom *Biljana* koja je preuzela podatke iz prethodno napravljenog direktorija *pig_biljana* te kreiranje stupaca *broj_ponavljanja* i *rijec*.

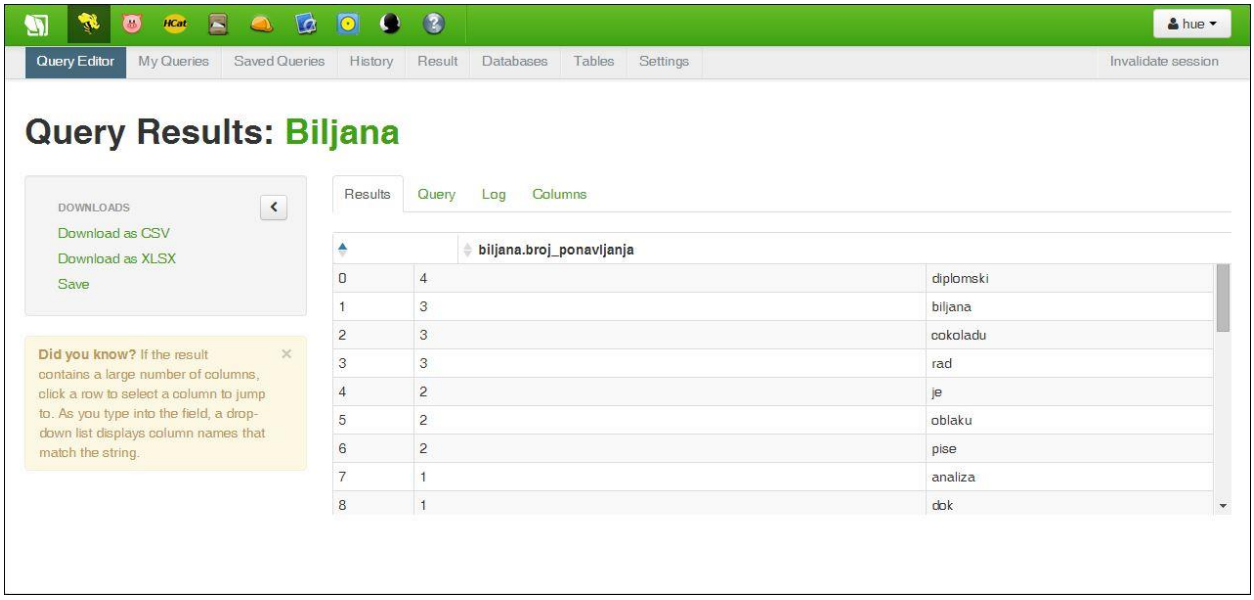


Slika 4.5. Kreiranje tablice u *HCatalog-u*

Prebacivanjem na *Hive*, koji je jednostavno rečeno program za uređivanje *SQL* upita nad velikim podacima, pojavljuje se uređivač za upisivanje *SQL* upita. Na s*lici 4.6* prikazan je jednostavan *SQL* upit koji izdvaja sve unose unutar tablice ***Biljana*** te sortira stupac ***broj_ponavljanja*** od najvećeg broja ponavljanja prema najmanjem, a stupac ***rijec*** sortira prema abecednom poretku.

Slika 4.6. *Hive – SQL* uređivač nad velikim podacima

Rezultati prethodnog *SQL* upita mogu se vidjeti na *slici 4.7.* Riječ ***diplomski*** pojavljuje se 4 puta u tekstu, dok riječi ***biljana, cokoladu*** i ***rad*** pojavljuju se tri puta, a ***je, oblaku, pise*** dva puta.



Slika 4.7. Rezultati *SQL* upita

S obzirom da je korišteni tekst kratak, za očekivati je mali broj pojavljivanja riječi, ali uz tekstualnu datoteku od nekoliko GB, korisnik će vrlo brzo moći uočiti pravu snagu obrade velikih skupova podataka u *Hadoop-u*.

4.2. Vizualizacija *clickstream* podataka

Clickstream podaci su statistički podaci koji prikazuju put kretanja i ponašanje posjetitelja na internet stranicama [38]. Internetske stranice korištenjem privremenih „kolačića“ (*engl. cookies*) prate aktivnost svakog korisnika koja se izračunava brojem pojedinog klika na mišu. Privremeni „kolačić“ ne određuje korisnika osobno, nego označava računalo jedinstvenom oznakom koja istječe onog trenutka kada korisnik ugasi pretraživač. Najčešći razlog prikupljanja i analize takvih podataka, kao što je objašnjeno u drugom poglavlju, je komercijalna iskoristivost, a ono što pružatelje usluga zanima jest:

- Tko je korisnik neke stranice, kada dolazi i što traži (kupuje)?
- Koliko se korisnik dugo zadržava te koji je najefikasniji put od pretrage do kupnje proizvoda?
- Koje proizvode korisnik najčešće kupuje zajedno i je li moguće korisniku sugerirati nekakav novi proizvod u stvarnom vremenu?
- Gdje je potrebno uložiti resurse kako bi se popravilo ili poboljšalo korisničko iskustvo na internet stranici pružatelja usluga?

Može se zaključiti da su *clickstream* podaci zapravo proces razmjene informacija između mrežnog korisnika i mrežnog poslužitelja. Takav proces zapisuje se u posebnim *log datotekama* koje u sebi sadrže podatke poput:

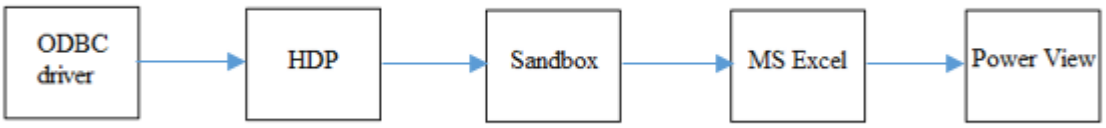
- Datuma i vremena zahtjeva/odgovora mrežnog poslužitelja
- Identifikacije posjetitelja - klijentova *IP* adresa
- Identiteta – na stranicama koje imaju implementiran oblik sigurnosne autentifikacije
- Zahtjeva i upita, obavljenih akcija i slično.

Što gledati i analizirati iz *log datoteka* ovisi o specifičnim poslovnim potrebama tvrtke? Ako se takve datoteke neprestano obogaćuju novim podacima „o ponašanju kupca“ pri sljedećem povratku na internet stranicu, tvrtke mogu predvidjeti njegovo ponašanje. Zajednički cilj svih koji žele uspjeti na e-tržištu je aktivno i kontinuirano usavršavanje ponude i proizvoda. Stoga je za očekivati da će se alati koji podržavaju ovakve analize razvijati do razine personaliziranog marketinga i usluga. Sljedeći primjer prikazuje jednu takvu analizu podataka iz *log datoteka*.

4.2.1. Potrebni alati za vizualizaciju *clickstream* podataka

U ovom primjeru potrebno je, kao što je prikazano na *slici 4.8*, redom koristiti navedene alate i upute dane u literaturi:

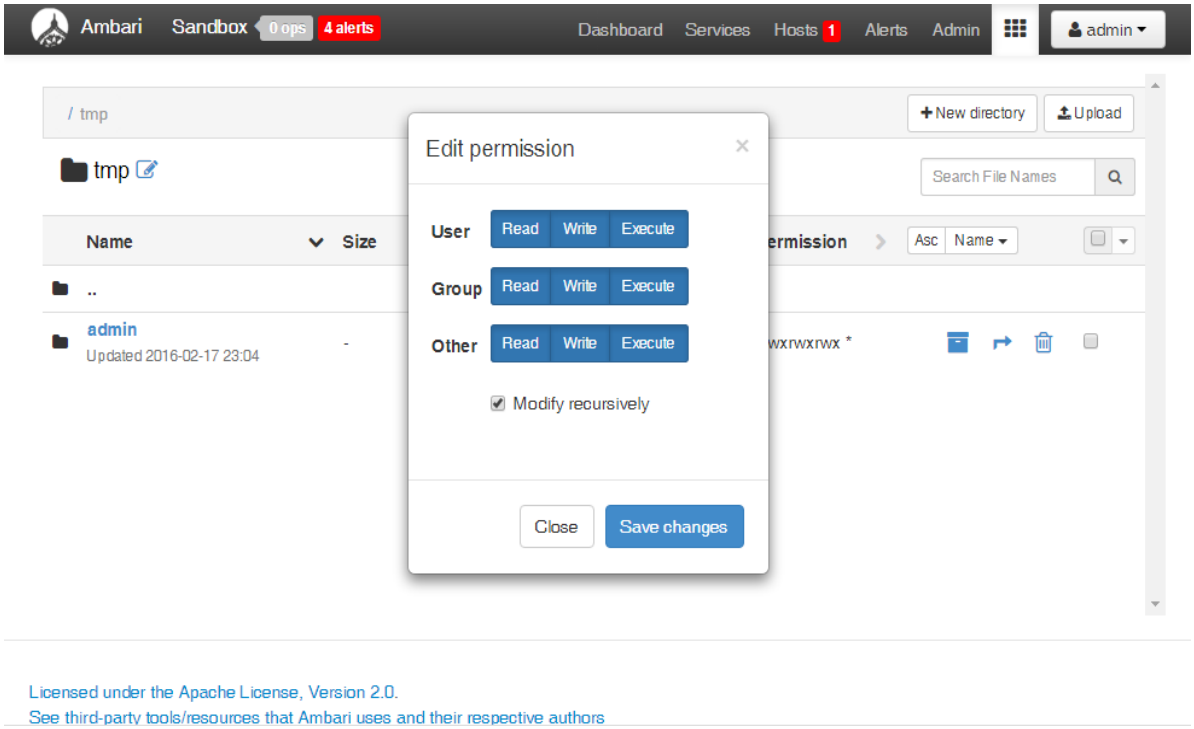
- Instalirati i konfigurirati *Hortonworks ODBC (Open Database Connectivity) driver* [39]
- Preuzeti, pohraniti i preraditi podatke u *Hortonworks Sandbox-u* [40]
- Pristupiti podacima pomoću *Microsoft Excel 2013 Professional Plus 64-bit*
- Vizualizirati dobivene podatke pomoću *Excel Power View*



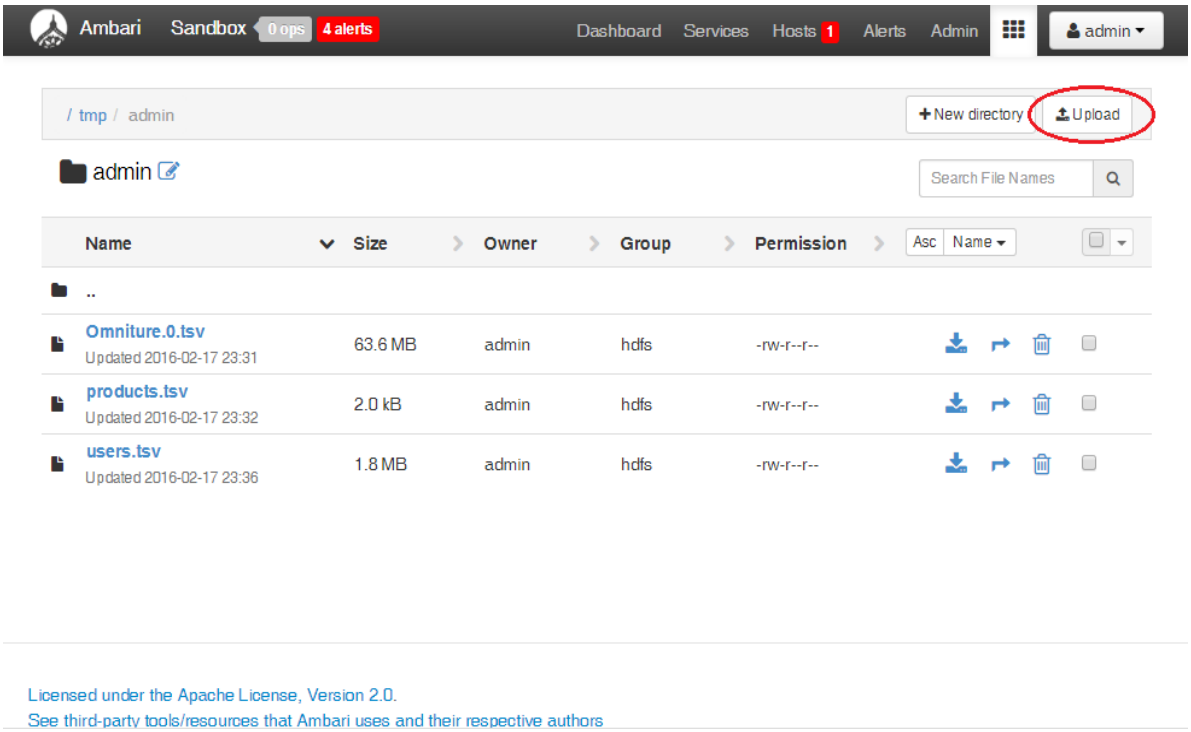
Slika 4.8. Slijed korištenja potrebnih alata za realizaciju drugog primjera

4.2.2. Priprema i filtriranje podataka

Kako bi se preuzete podatke moglo pohraniti u *Sandbox*, potrebno je u izborniku *Ambari* (gornji desni kut) izabrati *HDFS Files* i unutar *tmp direktorija* kreirati novi direktorij po imenu *admin* te desnim klikom omogućiti sve dozvole *(read, write, execute)* što se može vidjeti na s*lici 4.9*. Prema *slici 4.10*, unutar *admin* direktorija, postavljaju se (*engl. upload*) prethodno preuzete datoteke s osobnog računala, a to su *Omniture.0.tsv*, *users.tsv* te *products.tsv*.

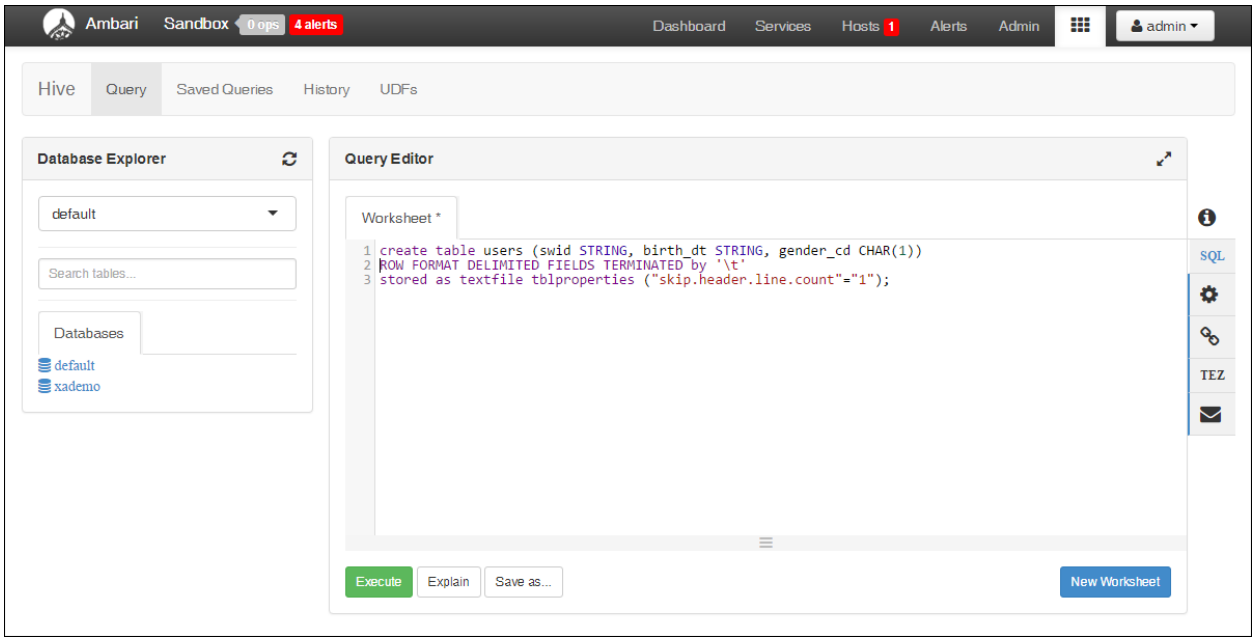


Slika 4.9. Kreiranje novog direktorija unutar *HDFS-a*



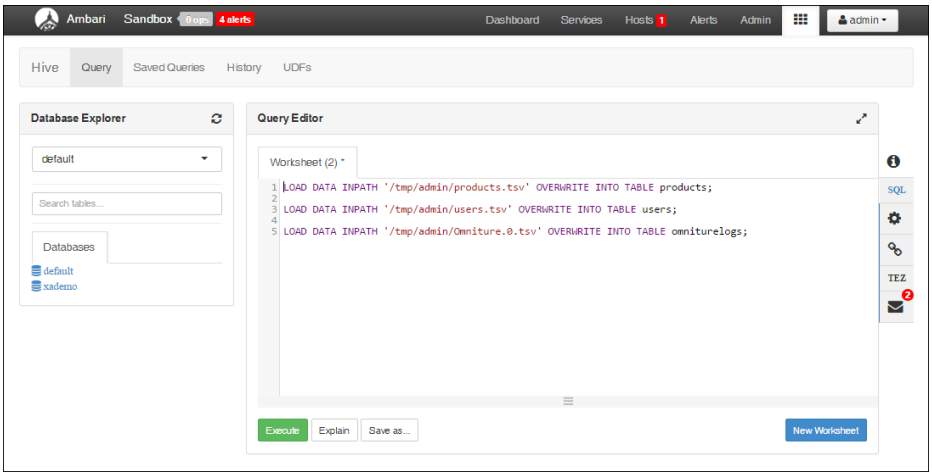
Slika 4.10. Postavljanje datoteka unutar admin direktorija

Nakon toga slijedi kreiranje tablica unutar *Hive-a* koje ćemo popuniti podacima koji su spremljeni u direktoriju *admin*. Jednostavnim *SQL* naredbama, posebno se kreiraju tablice *users*, *products* i *omniturelogs*, a na *slici 4.11* prikazano je kreiranje tablice *users*.



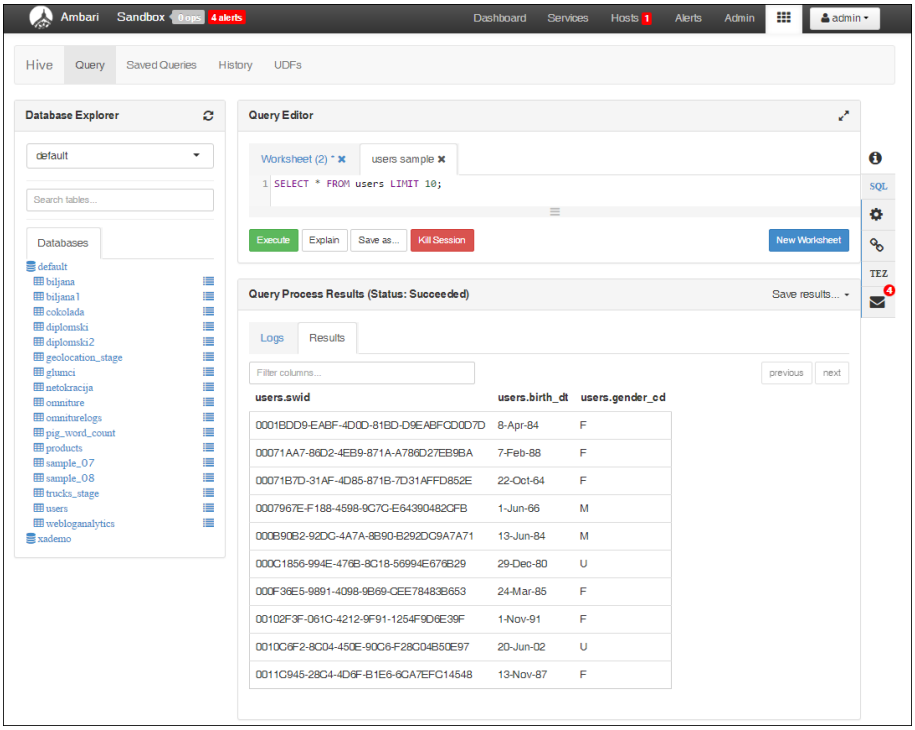
Slika 4.11. Kreiranje tablice *users* unutar *Hive-a*

Popunjavanje tablica s prethodno spremljenim podacima unutar *HDFS-a* izvršava se jednostavnim upitima (*engl. queries*) što se može vidjeti na *slici 4.12*.



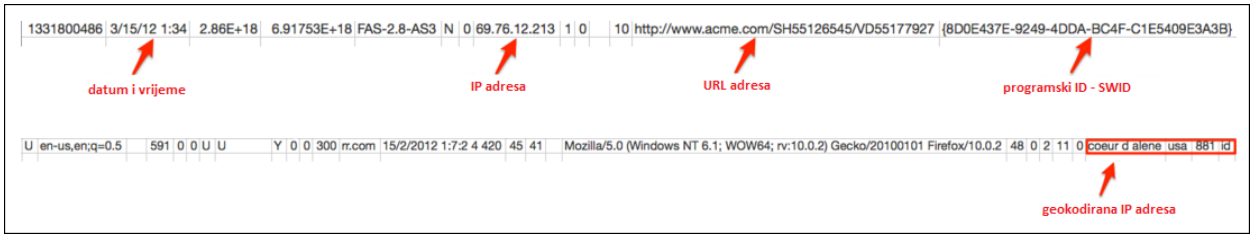
Slika 4.12. Popunjavanje tablica podacima spremljenim u *HDFS*

Kako bi se provjerilo jesu li upiti sa *slike 4.12* uspješno izvršeni, koristi se naredba *SELECT * FROM USERS*, gdje zvjezdica (*) označava „sve podatke unutar zadane tablice“, a zadana tablica je tablica *users*. Rezultati se mogu vidjeti na *slici 4.13*, a podaci koje tablica *users* sadrži su programski id korisnika (*engl. software id*), datum rođenja i spol.



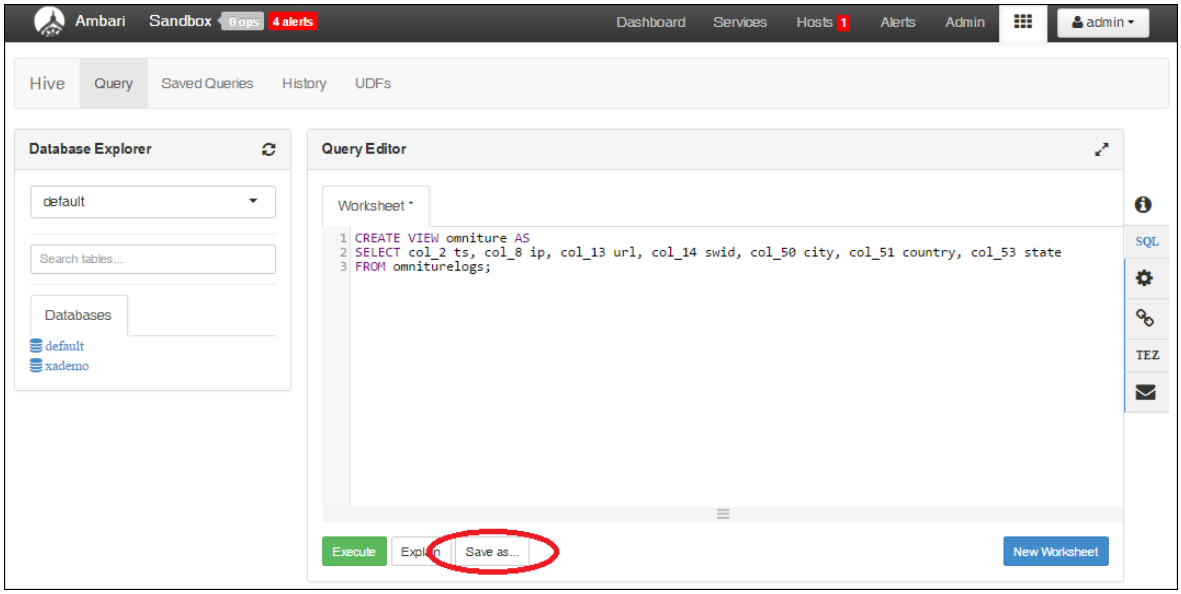
Slika 4.13. Prikaz prvih 10 redaka unutar tablice *users*

Može se primjetiti da tablica *omniturelogs* sadrži puno podataka koji se mogu koristiti u razne svrhe, a neki od njih su prikazani na *slici 4.14.* Najznačajniji podaci za ovaj primjer su datum i vrijeme, *IP* adresa, *URL* adresa, programski *ID*, grad, država, županija/država (*engl. state*).



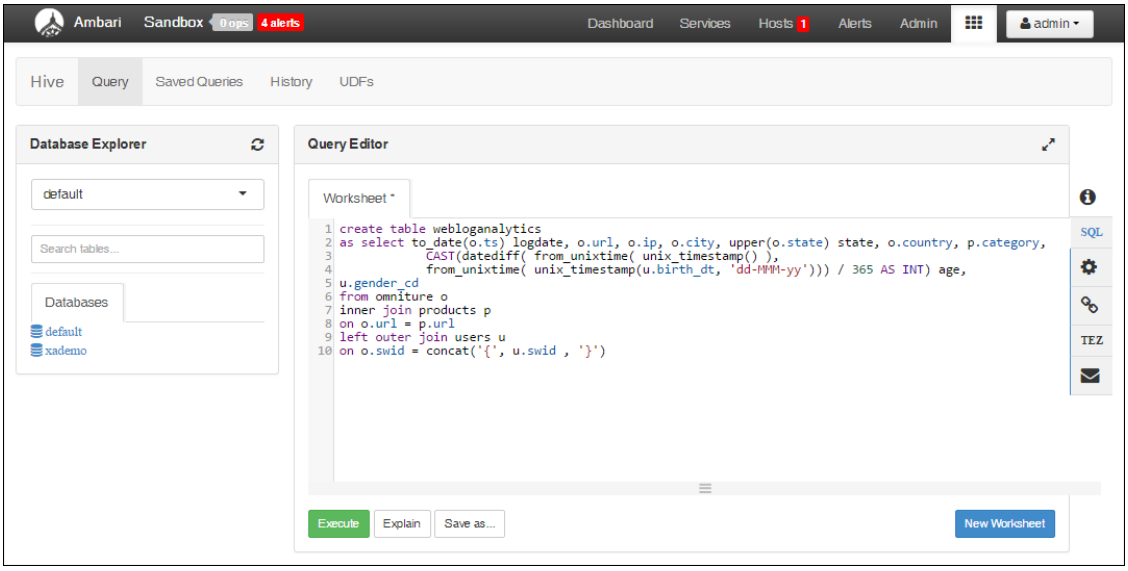
Slika 4.14. Podaci unutar *omniturelogs* tablice

Za lakše rukovanje podacima, može se napraviti posebna tablica iz izvorne tablice koja će sadržavati samo one podatke koji će se koristiti u daljnjoj analizi. Postupak kreiranja nove tablice prikazan je na *slici 4.15.* Nakon izvršavanja *SQL* upita, potrebno je novu tablicu spremiti klikom na *Save as.*



Slika 4.15. Nova tablica – *omniture*

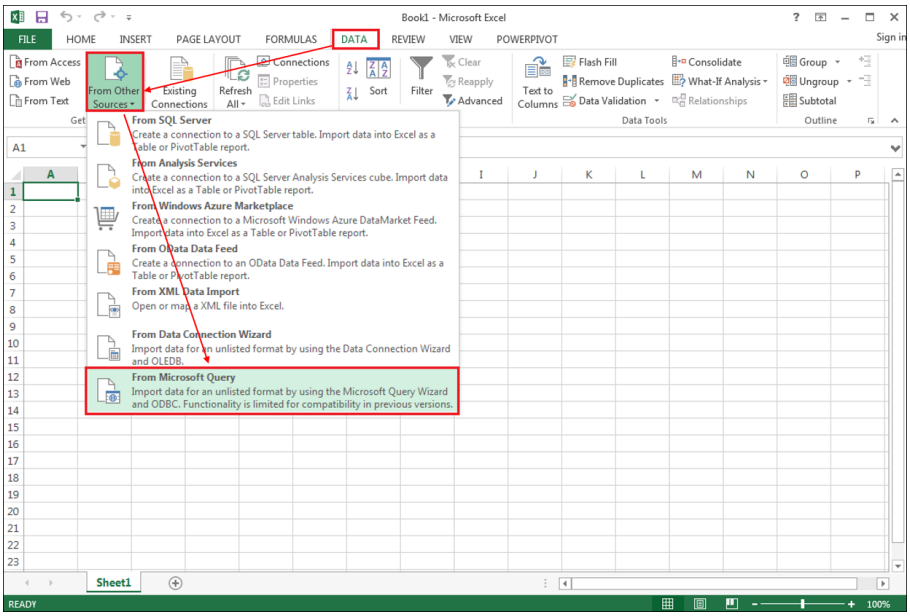
Zadnji korak u pripremi, odnosno filtriranju podataka je spajanje podataka iz triju tablica u jedinstvenu tablicu, prikazano na *slici 4.16*. Nova tablica, imena *webloganalytics*, sadržavat će podatke iz tablica *users*, *products* i *omniture*, a kako bi se izbjeglo ponavljanje istog podatka, „preklopit će“ se *URL* adrese iz tablica *omniture i products* te *programski id - swid* iz tablica *omniture i users*. Uspješnim kreiranjem jedinstvene tablice završen je postupak pripreme i filtriranja podataka. Nakon toga slijedi vizualizacija i analiza podataka.



Slika 4.16. Kreiranje jedinstvene tablice *webloganalytics*

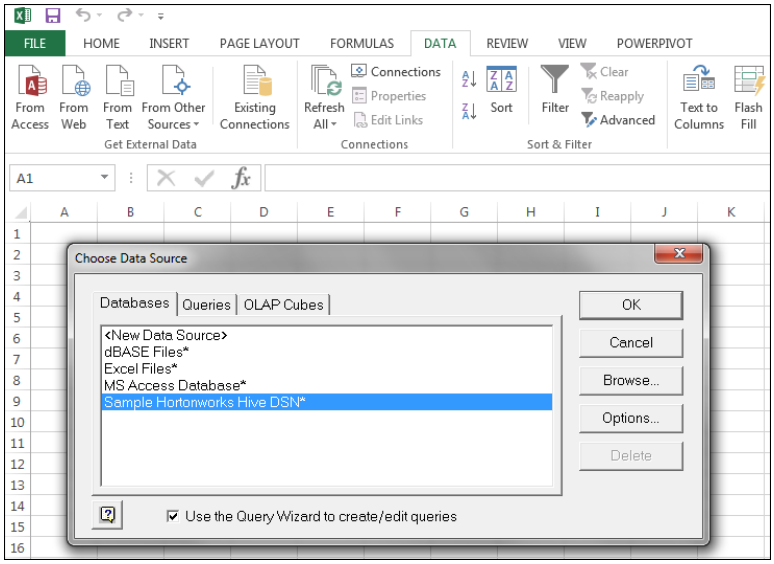
4.2.3. Analiza i vizualizacija podataka

Za analizu i vizualizaciju podataka koristit će se *Microsoft Excel Professional Plus 2013*. Iako je moguće koristiti i starije inačice, inačica 2013 je odabrana zbog elementa *Power View* koji omogućuje jednostavnu i efektnu vizualizaciju raznim dijagramima, mapama i slično. Kako bi se moglo pristupiti prethodno filtriranim podacima iz *Sandbox-a*, potrebno je povezati podatke s *Microsoft Excel-om* preko *ODBC* upravljačkog programa, što prikazuje *slika 4.17*.



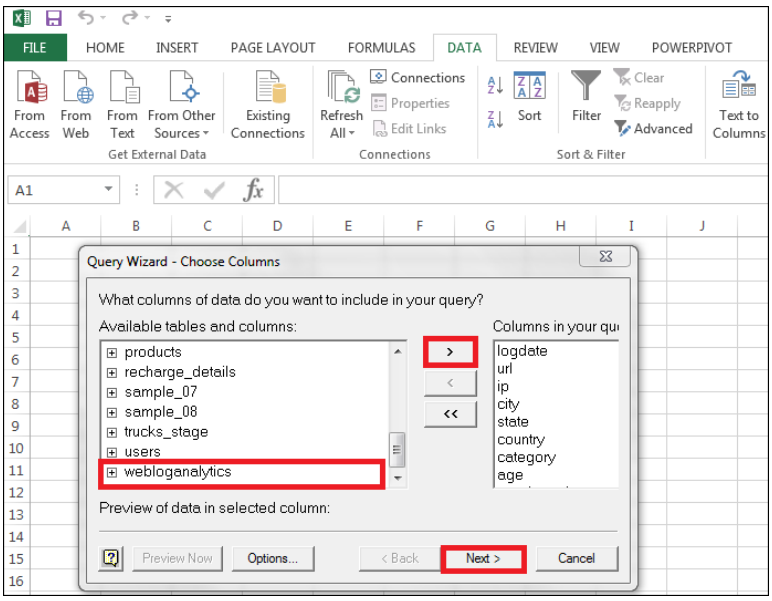
Slika 4.17. Povezivanje *MS Excel-a* s *ODBC* upravljačkim programom

Pojavom iskočnog prozora odabire se izvor podataka, a u ovom slučaju to je *Hortonworks Hive* kao jedan od ponuđenih opcija. Klikom na gumb *„OK“*, uspostavlja se veza između *Hortonworks HDP-a i MS Excel*-a, kako je i prikazano na s*lici 4.18.*



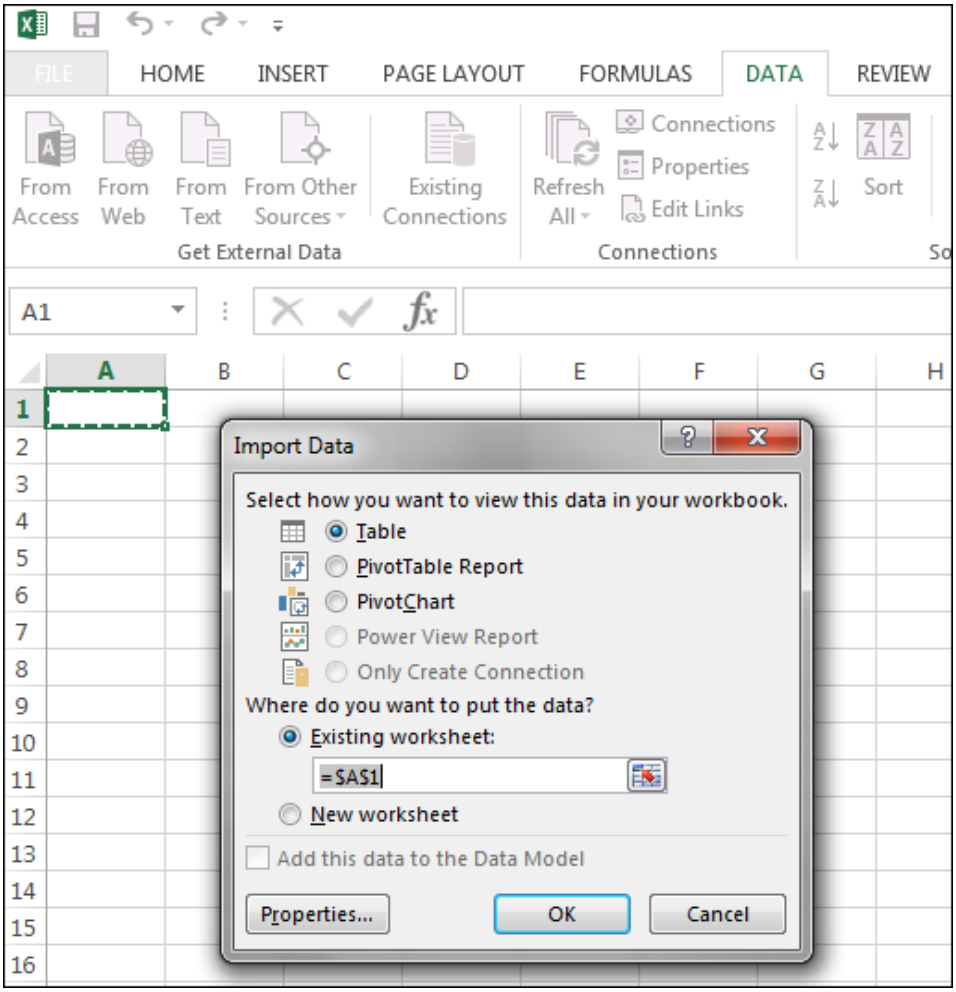
Slika 4.18. Odabir izvora podataka

Uspješnom uspostavom veze sa *Sandbox-om*, pojavljuje se asistent za upite tablica (*engl. query wizard*). Prema s*lici 4.19*, potrebno je odabrati prethodno kreiranu tablicu *webloganalytics* te kliknuti na desnu strelicu za dodavanje svih stupaca tablice. Za prelazak na idući korak, potrebno je kliknuti *„Next“*.



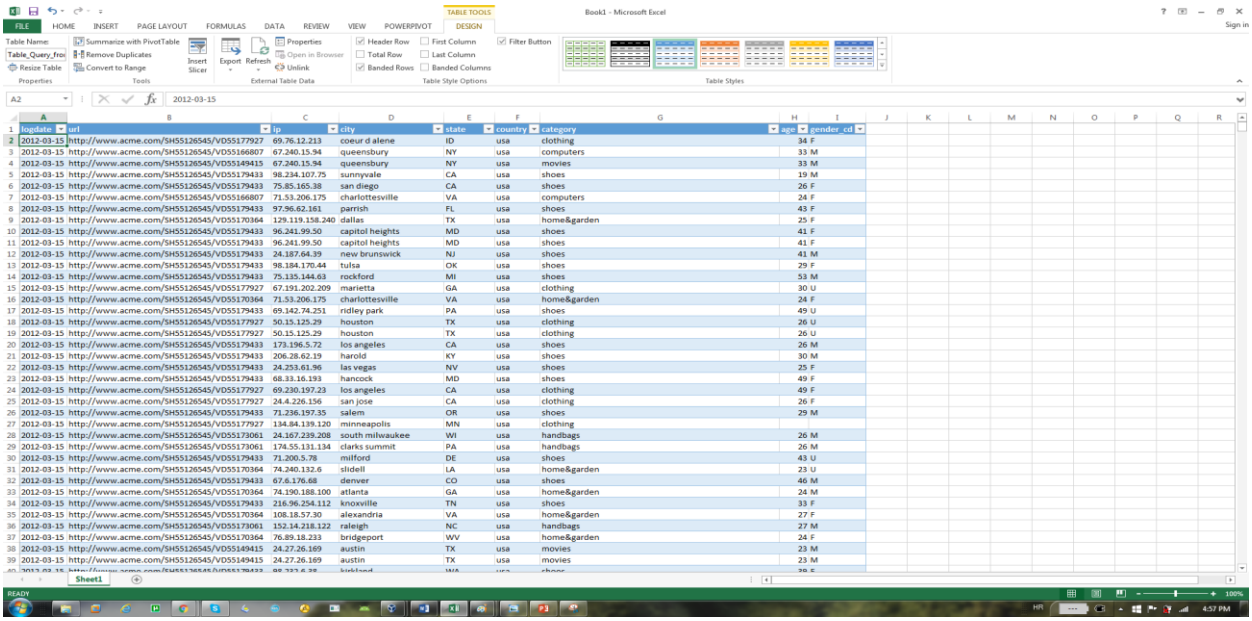
Slika 4.19. Odabir tablice podataka za analizu

U nastavku se pojavljuje prozor za filtriranje podataka *„Filter Data“* gdje je potrebno kliknuti gumb *„Next“*. Time se prelazi na novi prozor za sortiranje podataka gdje također ništa nije potrebno promijeniti te kliknuti gumb *„Next“*. Na posljednjem prozoru, klikom na gumb *„Finish“* podaci se prenose iz *Sandbox-a* i unose u *MS Excel*. *Slika 4.20* prikazuje opcije uvoza podataka (*engl. import*) kao što je tablica, izvještaj ili graf. Odabirom prvo ponuđene opcije „tablica“ i klikom na gumb *„OK“* završava se proces spajanja i uvoza podataka u *MS Excel.*



Slika 4.20. Mogućnosti načina uvoza podataka u MS Excel

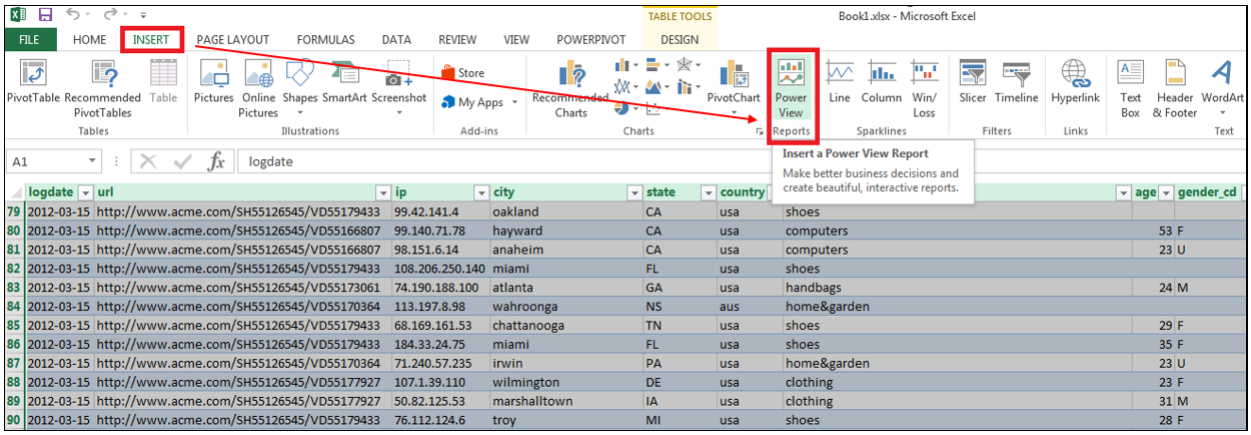
Nakon uspješno završenog procesa spajanja i uvoza podataka, stvara se prvi radni list u radnoj knjizi koji sadrži podatke tablice *webloganalytics*. U tablici se nalaze stupci: datum, *URL* adresa, *IP* adresa, grad, država/županija, država, kategorija proizvoda, dob i spol korisnika što prikazuje slika *4.21.*



Slika 4.21. *Webloganalytics* tablica unutar *MS Excel-a*

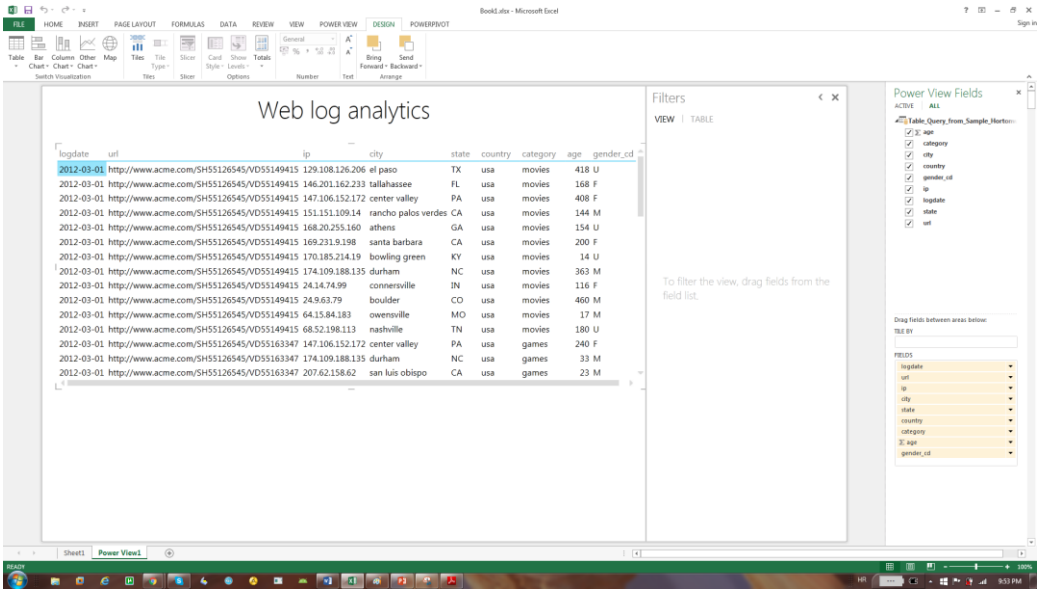
Vizualizacija podataka pomaže pri optimizaciji internet stranica te pretvaranju većeg broja posjeta korisnika u prodaju, odnosno prihod. Drugim riječima, ukoliko poduzeće iz vizualiziranih podataka iz *log datoteka* dobije uvid da se korisnici ne vraćaju na njihovu stranicu kako bi kupili nekakav proizvod, tada pristupaju optimizaciji stranice kako bi privukli kupce. Jednako tako, ako se kupci rado vraćaju na određenu stranicu, poduzeće profitira, jer se širi interesna grupa, povećava prodaja, a ujedno i prihodi.

Jedan od najlakših *BI* (*engl. Business Intelligence*) aplikacija za pristup i analizu podataka je navedeni *MS Excel*, odnosno njegov element *Power View* koji dolazi u verziji *Professional Plus 2013*. Prema *slici 4.22* prikazana je putanja za stvaranje novog *Power View* izvještaja.



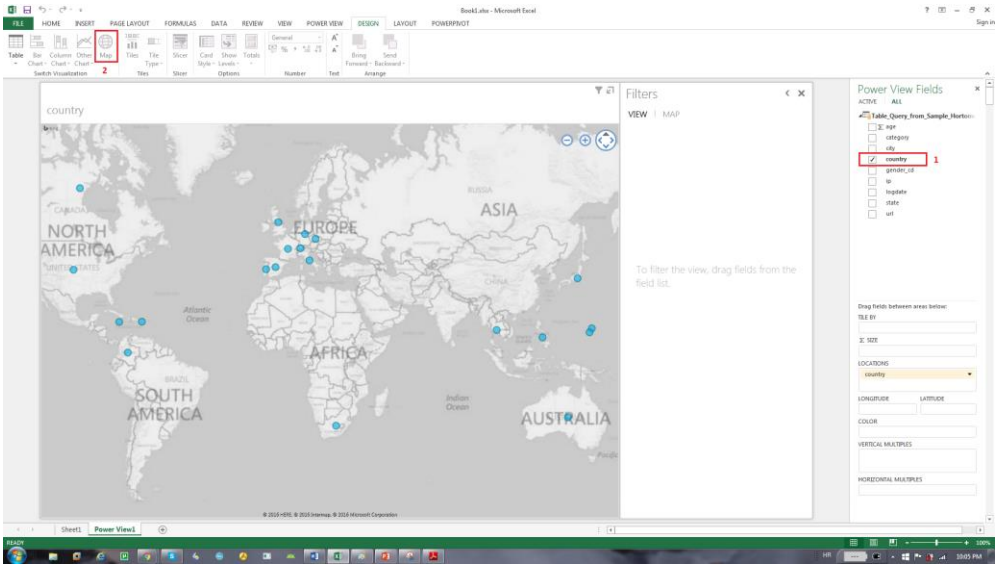
Slika 4.22. Stvaranje *Power View* izvještaja

Kreiranjem prvog *Power View* izvještaja stvara se novi radni list (izvještaj) prikazan na s*lici 4.23.* S lijeve strane prikazana je tablica s pripadajućim stupcima, pokraj se nalaze filteri te polja za daljnju obradu i analizu koja se mogu dodavati u okvirima za izbor (*engl. checkbox*).



Slika 4.23. Početni izgled *Power View* izvještaja

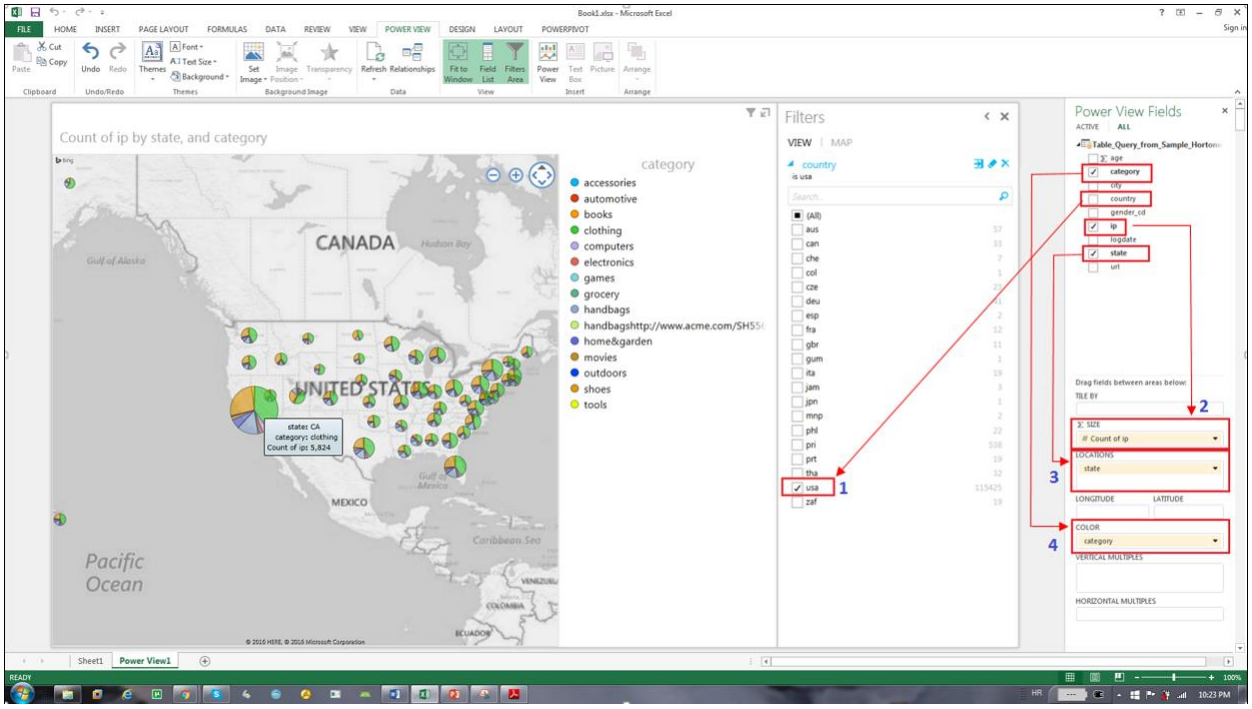
Pod pretpostavkom da želimo prikazati *clickstream podatke* određene internet stranice po lokacijama na globalnoj razini, označit ćemo samo polje *country* (ostale ćemo isključiti) te u Dizajn izborniku izabrati *Map*, što je prikazano na s*lici 4.24.*



Slika 4.24. Prikaz korisničkih posjeta na globalnoj razini

Dodatno, proizvoljno filtriranje podataka, vidljivo na *slici 4.25*, prikazuje ukupan broj pogledanih/kupljenih proizvoda po kategorijama (*modni dodaci, automobilska oprema, knjige, odjeća, računala, elektronika, igre, hrana, torbice, kuća&vrt, filmovi, oprema za izvan kuće, cipele i alati*) u različitoj boji za svaku saveznu državu unutar SAD-a zasebno. Može se primjetiti da je u saveznoj državi Kalifornija, najviše prodanih artikala (5824) u kategoriji *odjeća* (*engl. clothing*). Postupak analize, odnosno filtriranja podataka je slijedeći:

1. Za prikaz podataka pojedine države, odaberemo polje *country* te ga mišem povučemo u polje *Filters* i u okvirima za izbor, odaberemo željenu državu (izborom države USA, prikazat će se samo podaci za USA, a razlog baš tog odabira je testirano najveći prikaz podataka prema saveznim državama)
2. Prebacivanjem polja *ip* u okvir *SIZE*, prikazat će se ukupan broj posjećenih *IP* adresa za određenu skupinu proizvoda
3. Pridruživanjem polja *state* u okvir *LOCATIONS*, prikazat će se podaci za pojedinu saveznu državu unutar SAD-a
4. Kako bi se prikazale kategorije *products* (kupljeni proizvodi*)* u boji na trenutnoj mapi, potrebno je polje *products* pridružiti u okvir *CATEGORY*

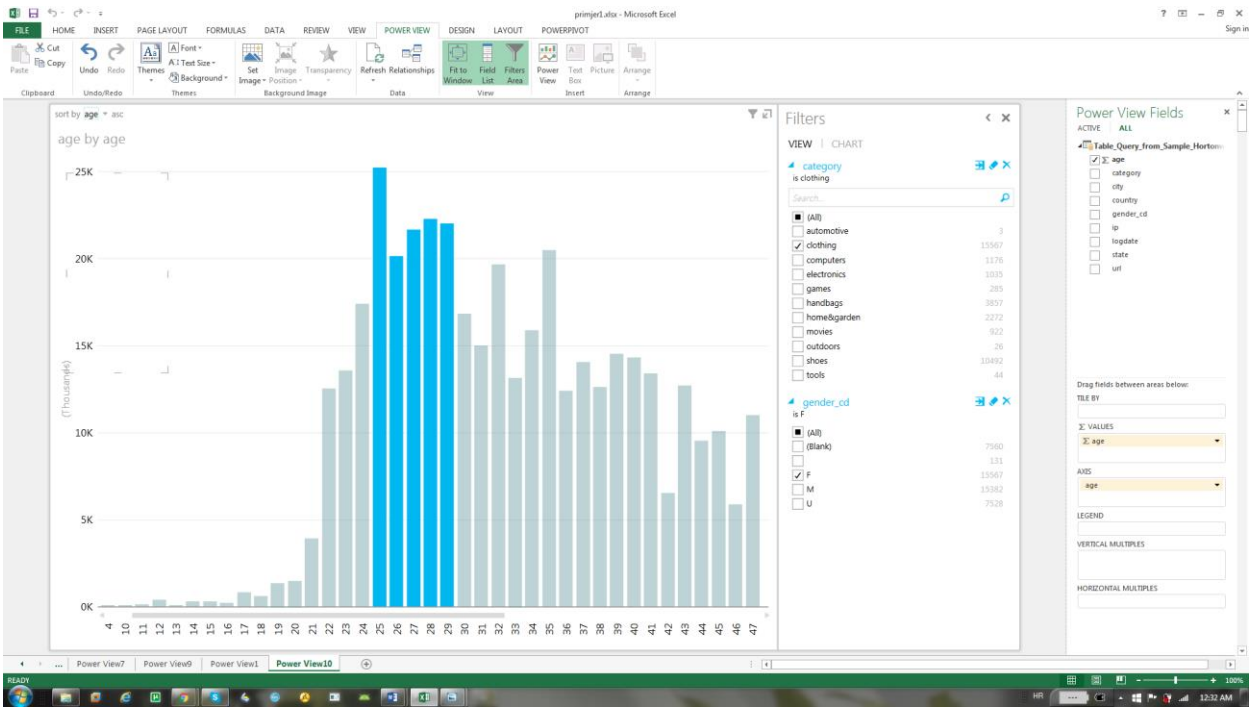


Slika 4.25. Prikaz kupljenih proizvoda po kategorijama u pojedinoj državi

Na osnovu prethodno dobivenih podataka, klijent (tvrtka) može zatražiti statistiku podataka kupljene odjeće po dobi i spolu kako bi na osnovu rezultata dodatno optimizirao svoju internet stranicu i ponudu na njoj. Kako bi se to napravilo, potrebno je kreirati novi *Power View* izvještaj preko *Insert/Power View* te dodati nove filtere:

1. U *Power View Fields* području izabrati polja *ip* i *age*
2. Povući mišem polje *category* iz područja *Power View Fields* u područje *Filters*
3. Ponoviti postupak za polje *gender* te izabrati F-žene (*engl. female*)
4. Odabrati stvaranje grafa preko *File/Column Chart/Clustered Column*
5. Povući mišem polje *age* u *AXIS* okvir te iz njega izbrisati polje *ip*

Dobiveni graf na s*lici 4.26* prikazuje da je većina žena koje su kupile odjeću na klijentovoj stranici u dobi između 25 i 29 godina. Pomoću te informacije, klijent može dodatno optimizirati svoju internet stranicu za taj segment tržišta te istražiti na koji način privući nove generacije kupaca i djelovati u tom smjeru.

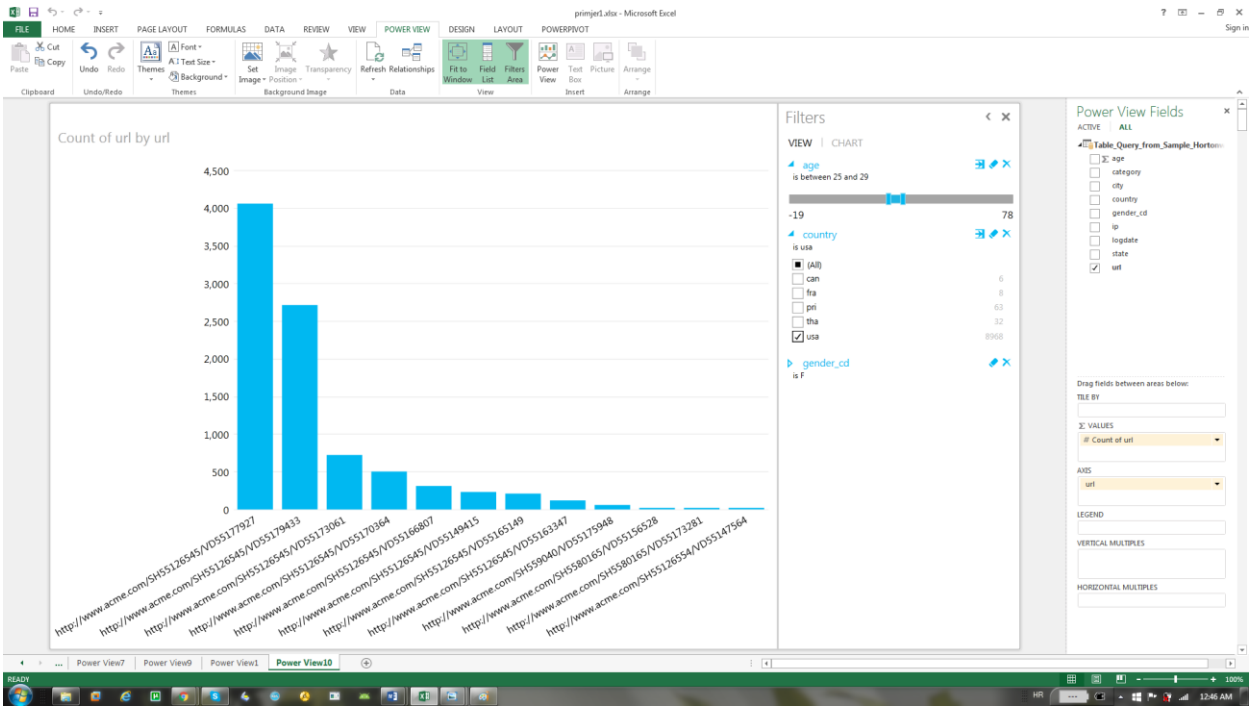


Slika 4.26. Prikaz statističkih podataka za kupljenu odjeću po dobi i spolu

Pretpostavimo da prethodno dobiveni podaci sadrže informacije o internet stranicama s visokom stopom napuštanja (*engl. bounce rate*). *Bounce rate* predstavlja postotak posjetitelja koji započinju pretraživanje na jednoj određenoj internet stranici, a zatim ju naglo napuštaju (*engl. bounce*) bez detaljnijeg pretraživanja po ponuđenim linkovima i sadržajima [41]. Stopa

napuštanja je mjera učinkovitosti internet stranice u poticanju posjetitelja da nastavi sa svojim pretraživanjem. Stranica s visokom stopom napuštanja je ona stranica koju je korisnik napustio odmah nakon učitavanja, bez dodatnih pretraživanja.

Filtrirajući *URL* podatke (povući mišem polje *url* u *AXIS* okvir te iz njega izbrisati polje *age*) prethodno dobivene dobne skupine (žene između 25 i 29 godina), mogu se dobiti točni podaci koje web stranice je potrebno dodatno optimizirati. Pogledom na graf na s*lici 4.27*, može se uočiti da postoje dvije web stranice s visokom stopom napuštanja.



Slika 4.27. Grafički prikaz*„Bounce rate-a“* internet stranica

Prema tim podacima, klijent može redizajnirati navedene web stranice te testirati novi dizajn na temelju ciljane dobne skupine, smanjiti stopu napuštanja, stimulirati zadržavanje kupca na web stranici i povećati prodaju proizvoda.

Ovim primjerom prikazano je kako se iz naizgled nelogičnih i nerazumljivih podataka može doći do vrijednih informacija za razvoj tvrtke i uvida trenutnog stanja na tržištu. Također, bitno je napomenuti da je za iskoristivost dobivenih zaključaka potrebna pravilna analiza i interpretacija koja nije moguća bez stručnog kadra. Svrha primjera je pokazati maleni dio spektra mogućnosti koje se mogu realizirati na platformama za obradu velikih skupova podataka.

5. ZAKLJUČAK

U ovom radu prikazana je kratka teorijska podloga o tome što su veliki skupovi podataka u oblaku računala i kako se mogu iskoristiti. Objašnjeni su programski alati i platforme koje su se koristile u izradi praktičnog dijela rada. Cilj rada bio je istražiti zahtjeve i prikazati mogućnosti velikih skupova podataka te pomoću primjera predstaviti najjednostavnije načelo i svrhu korištenja u stvarnom životu. Za potrebe rada korištena je besplatna *Hortonworks* podatkovna platforma (HDP) koja u programskom okruženju *Hadoop* sadrži razne alate. Sve manipulacije odvijale su se preko grafičkog sučelja *Hue* gdje je korišten *HDFS sustav* za pohranu i učitavanje podataka. *HCatalog* je služio za kreiranje potrebnih tablica dok su se podaci iz tih tablica obrađivali preko sustava *Hive*. Korišten je i skriptni jezik *Pig Latin* te *MS Excel* i njegov alat *Power View* za vizualizaciju podataka.

Prvi primjer u ovom diplomskom radu prikazuje obradu tekstualne datoteke kroz skriptni jezik *Pig Latin* te način i brzinu prebrojavanja riječi u velikim skupovima podataka. U drugom primjeru korištena je otvorena datoteka zapisa s internet stranica iz koje se analizom i vizualizacijom podataka došlo do novih, neočekivanih informacija koje se mogu iskoristiti za poboljšanje poslovanja zamišljene tvrtke. Iz drugog primjera se može zaključiti da koliko god se neki podaci čine nelogičnima i neupotrebljivima, oni ipak u sebi sadrže skrivenu vrijednost ako im se pristupi na prikladan način. To su već prepoznale vladine organizacije, vojna industrija, robotika i medicina, a u budućnosti se mogu očekivati primjene u svemirskim tehnologijama, poljoprivredi i kućanstvu.

LITERATURA

[1] P. Mell, T. Grance,The NIST Definition of Cloud Computnig, http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf, pristup: 09.10.2015.

[2] G. Press, A Very Short History of Big Data http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/, pristup: 11.10.2015.

[3] D. de Solla Price, „Science since Babylon“, Yale University Press, New Haven, Conn., 1961.

[4] H. Becker, „Can Users Really Absorb Data at Today's Rates? Tomorrow's?“, Data Communications, 1986.

[5] P.Lyman, H.R.Varian, „How Much Information?“, http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf, pristup: 10.10.2015.

[6] D.Laney, „3D Data Management: Controlling Data Volume, Velocity and Variety“, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, pristup: 10.10.2015.

[7] D. Boyd, K. Crawford, „Critical Questions for Big Data“, http://www.tandfonline.com/doi/pdf/10.1080/1369118X.2012.678878, pristup: 10.10.2015.

[8] V. Cipan, Big Data: Veliki izazovi, velike prilike, a možda i veliki „buzzword“ http://www.netokracija.com/big-data-54787, pristup: 11.10.2015.

[9] D. Marjanović, Podaci su zlato – šta je big data, šta su Hadoop i Spark, i kako krenuti sa njima?, http://startit.rs/big-data-hadoop-apache-spark/, pristup: 11.10.2015.

[10] J. Waite, A Simple Guide to Big Data, https://www.linkedin.com/pulse/20140912112657-9245190-a-simple-guide-to-big-data-2014, pristup: 12.10.2015.

[11] M. Biberović, Combisovci tvrde: Big Data izlazi iz eksperimentalne faze; tržištu sada trebaju stručnjaci, http://www.netokracija.com/big-data-combis-konferencija-husky-107636, pristup: 11.10.2015.

[12] Z.S., Zuboff, „In the Age of the Smart Machine: The Future and Power of Work“, New York, 1988.

[13] D. J. Patill, Building data science team, http://radar.oreilly.com/2011/09/building-data-science-teams.html, pristup: 11.10.2015.

[14] P. C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, G. Lapisa, „Understanding Big Data“, The McGraw-Hill Companies, 2012

[15] SAS, Big Data; What it is and why it matters, http://www.sas.com/en_us/insights/big-data/what-is-big-data.html, pristup: 12.10.2015.

[16] R. M. Burton, D.Mastrangelo, F.Salvador, „Big Data and Organization Design“, Vol. 3, No.1 (2014), Journal of Organization Design, 2014.

[17] J. R. Galbraith, The Star Model, http://www.jaygalbraith.com/images/pdfs/StarModel.pdf, pristup: 13.10.2015.

[18] J. Lukić, Uticaj Big Data-e na strategiju, http://www.hadoop-srbija.com/uticaj-big-data-e-na-strategiju/, pristup:13.10.2015.

[19] Hortonworks Inc., Blog, February 25th, 2015, http://hortonworks.com/big-data-insights/industries-benefit-big-data/, pristup: 13.10.2015

[20] M. van Rijmenam, The Advantages and Disadvantages Of Real-Time Big Data Analytics, https://datafloq.com/read/the-power-of-real-time-big-data/225, pristup: 14.10.2015.

[21] T. Moore, The impact of Big Data on Society, Western Oregon University, http://www.wou.edu/~tmoore08/Portfolio%20Essay%203.pdf, pristup: 14.10.2015.

[22] D. Feinleib, The Big Data Landscape, http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/, pristup: 15.10.2015.

[23] Wikipedia, Apache Hadoop, https://en.wikipedia.org/wiki/Apache_Hadoop, pristup: 16.10.2015.

[24] SAS, Hadoop – What is it and why does it matter?, http://www.sas.com/en_my/insights/big-data/hadoop.html, pristup: 15.10.2015.

[25] What is open source?, https://opensource.com/resources/what-open-source, pristup: 15.10.2015.

[26] Apache Hadoop, What Is Apache Hadoop?, http://hadoop.apache.org/, pristup: 16.10.2015.

[27] Wikipedia, Apache Hadoop, https://en.wikipedia.org/wiki/Apache_Hadoop, pristup: 17.10.2015.

[28] W.T., White, „Hadoop, The Definitive Guide“, O'Reilly Media, 2009.

[29] A.S., Achari, „Hadoop Essentials“, PACKT Publishing, Birmingham – Mumbai, 2015.

[30] M.M., Mois, „Apache Hadoop Tutorial, The Ultimate Guide“, Web Code Geeks, Exelixis Media P.C., 2016

[31] Hortonworks Inc., Quick Facts, Last Updated: October 2015, http://hortonworks.com/about-us/quick-facts/, pristup 25.10.2015.

[32] http://hortonworks.com/products/hortonworks-sandbox/#install, pristup: 25.10.2015.

[33] Hortonworks Inc., Hortonworks Sandbox with VirtualBox, http://hortonworks.com/wp-content/uploads/2015/07/Import_on_Vbox_7_20_2015.pdf, pristup: 25.10.2015.

[34] Apache Ambari, Introduction, https://ambari.apache.org/, pristup 26.10.2015.

[35] Apache Hive, https://hive.apache.org/, pristup: 27.10.2015.

[36] Confluence, Hcatalog UsingHCat, https://cwiki.apache.org/confluence/display/Hive/HCatalog+UsingHCat, pristup: 30.10.2015.

[37] Hortonworks Inc., Apache Pig, http://hortonworks.com/hadoop/pig/#section_1, pristup: 05.11.2015.

[38] Skladistenje, autor L.I., Klikni i reći ću ti što želiš, http://www.skladistenje.com/klikni-i-reci-cu-ti-sto-zelis/, pristup: 10.11.2015.

[39] Hortonworks, http://hortonworks.com/downloads/, pristup: 11.11.2015.

[40] https://s3.amazonaws.com/hw-sandbox/tutorial8/RefineDemoData.zip, pristup: 11.11.2015.

[41] Wikipedia, Bounce rate, https://en.wikipedia.org/wiki/Bounce_rate, pristup: 15.11.2015.

SAŽETAK

Cilj ovog diplomskog rada bio je istražiti zahtjeve i mogućnosti obrade velikih skupova podataka u oblaku računala, prikazati gdje se veliki podaci koriste te na nekoliko primjera, predstaviti alate *Hadoop* ekosustava i njihovu primjenu. Pri upotrebi velikih skupova podataka ključna je analiza i pravilna interpretacija. Spajanjem starih i modernih tehnologija izgrađena je potrebna infrastruktura, kojom se može doći do novih, skrivenih informacija koje će pomoći u rješavanju modernih problema. Primjenom *Pig, Hive, HCatalog i MS Excel* alata na otvorenoj datoteci zapisa s internet stranica, izvršena je analiza i vizualizacija dobivenih podataka. Grafičkim prikazom dobivene su potrebne informacije za daljnju optimizaciju rada zamišljene tvrtke. Veliki podaci mogu se koristiti u svim područjima ljudskog djelovanja te se može zaključiti da će imati jednu od glavnih uloga pri razvoju novih informacijskih sustava u budućnosti.

Ključne riječi: analiza podataka, Hadoop, računarstvo u oblaku, veliki skupovi podataka, vizualizacija podataka.

ABSTRACT

The aim of this diploma thesis was to investigate the requirements and possibilities of processing large data sets in cloud computing; to see where big data is used and on the basis of a few examples, present tools of Hadoop ecosystem and its implementation. It is crucial to have correct analysis and interpretation when using large data sets. For that matter, necessary infrastructure is built combining old and modern technologies, which can lead to new, hidden information that will help solve modern problems. Applying *Pig, Hive, HCatalog and MS Excel* tools on the downloaded data log files, an analysis and visualization of data was obtained. Graphic presentation gave the necessary information for further optimization of imaginary company. Big Data can be used in all areas of human activity and it can be concluded that it will have a major role in the development of new information systems in the future.

Key words: Data Analysis, Hadoop, Cloud Computing, Big Data, Data Visualization.

ŽIVOTOPIS

Biljana Krpan, rođena je 14.07.1991. godine u Đakovu, Republika Hrvatska. Osnovnu školu „Ivan Goran Kovačić“ završava 2006. godine u Đakovu te iste godine završava Osnovnu glazbenu školu pri OŠ „Ivan Goran Kovačić“ u Đakovu. Srednju školu, Gimnazija „A.G.Matoš“ – prirodoslovno-matematički odjel, upisuje 2006. godine. Državnu maturu polaže 2010. godine te iste godine upisuje sveučilišni Preddiplomski studij Računarstva na Elektrotehničkom fakultetu u Osijeku. U ljeto 2013. godine odlazi na prvu međunarodnu IAESTE stručnu studentsku praksu u Bangkok, Tajland gdje 3 mjeseca radi na King Mongkut's University of Technology North Bangkok. U rujnu 2013. godine završava Preddiplomski studij Računarstva u Osijeku s najvišim pohvalama (Summa Cum Laude) te upisuje sveučilišni Diplomski studij Procesnog računarstva na Elektrotehničkom fakultetu u Osijeku. U ljeto 2014. godine odlazi na drugu međunarodnu IAESTE stručnu studentsku praksu u Seoul, Južna Koreja, gdje 3 mjeseca radi u državnoj agenciji za internet sigurnost - Korea Internet & Security Agency.

Sudjelovala je na prvom Young ICT Leaders' Forum-u u prosincu 2015. godine u Busanu, Južna Koreja, koji je organiziran pod vodstvom ITU-a (International Telecommunication Union) i NIA-e (National Information Society Agency). U svibnju 2016. sudjeluje na LEAP Summit-u u Zagrebu.

PRILOZI (CD)

Prilog 1. Pisana verzija diplomskog rada u *.doc* formatu

Prilog 2. Pisana verzija diplomskog rada *.pdf* formatu

Prilog 3. Microsoft Excel dokument sa analiziranim podacima