# Predgovor

Kako je ova knjiga nastala? Poreklo je duboko usađeno u naše putovanje od nauke o podacima do inženjerstva podataka. Često se u šali nazivamo *bivšim naučnicima podataka*. Obojica smo imali iskustvo da nam je dodeljen posao na projektima nauke o podacima, a onda smo se mučili da izvršimo te projekte zbog nedostatka odgovarajućih temelja. Naše putovanje u inženjerstvo podataka počelo je kada smo preduzeli zadatke inženjerstva podataka kako bismo izgradili temelje i infrastrukturu.

Sa razvojem nauke o podacima, kompanije su rasipnički trošile na talente iz nauke o podacima, nadajući se bogatim nagradama. Često, naučnici podataka su se borili sa osnovnim problemima koje njihovo obrazovanje i obuka nisu rešavali – prikupljanje podataka, čišćenje podataka, pristup podacima, transformacija podataka i infrastruktura podataka. To su problemi koje inženjerstvo podataka ima za cilj da reši.

## Šta ova knjiga nije

Pre nego što objasnimo o čemu je ova knjiga i šta ćete iz nje dobiti, brzo ćemo preći šta ova knjiga *nije*. Ova knjiga nije o inženjerstvu podataka koristeći određeni alat, tehnologiju ili platformu. Iako mnoge odlične knjige prilaze tehnologijama inženjerstva podataka iz ove perspektive, te knjige imaju kratak vek trajanja. Umesto toga, fokusiramo se na fundamentalne koncepte iza inženjerstva podataka.

## O čemu je ova knjiga

Ova knjiga ima za cilj da popuni prazninu u trenutnom sadržaju i materijalima inženjerstva podataka. Iako nema nedostatka tehničkih resursa koji se bave specifičnim alatima i tehnologijama inženjerstva podataka, ljudi se bore da razumeju kako da sklope ove komponente u koherentnu celinu koja se primenjuje u stvarnom svetu. Ova knjiga povezuje tačke celog životnog ciklusa podataka. Ona vam pokazuje kako da povežete različite tehnologije kako bi služile potrebama korisnika podataka kao što su analitičari, naučnici podataka i inženjeri mašinskog učenja. Ova knjiga funkcioniše kao dopuna knjigama O'Reilly koje pokrivaju detalje određenih tehnologija, platformi i programskih jezika.

Velika ideja ove knjige je *životni ciklus inženjerstva podataka*: generisanje podataka, skladištenje, unos, transformacija i servisiranje. Otkako postoje podaci, svedoci smo uspona i padova brojnih specifičnih tehnologija i proizvoda vendora, ali faze životnog ciklusa inženjerstva podataka su ostale suštinski nepromenjene. Sa ovim okvirom, čitalac će steći dobro razumevanje primene tehnologija na stvarne poslovne probleme.

Naš cilj ovde je da mapiramo principe koji dopiru kroz dve ose. Prvo, želimo da razdvojimo inženjerstvo podataka u principe koji mogu obuhvatiti *bilo koju relevantnu tehnologiju*. Drugo, želimo da predstavimo principe koji će izdržati test *vremena*. Nadamo se da ove ideje odražavaju lekcije naučene kroz preokret u tehnologiji podataka poslednjih dvadeset godina i da će naš mentalni okvir ostati koristan sledeću deceniju ili više u budućnosti.

Jedna stvar koju treba napomenuti: mi se neopravdano pridržavamo *oblak na prvom mestu* (claud-first) pristupa. Vidimo oblak kao fundamentalno transformativni razvoj koji će trajati decenijama; većina lokalnih sistema i radnih opterećenja podataka na kraju će se preseliti na klaud hostinge. Pretpostavljamo da su infrastruktura i sistemi *prolazni* i *skalabilni* i da će se inženjeri podataka oslanjati na raspoređivanje upravljanih usluga u oblakuu. Iako će većina koncepata u ovoj knjizi biti prevedena na ne-klaud okruženja.

## Kome je namenjena knjiga

Naša primarna ciljna publika za ovu knjigu se sastoji od tehničkih praktičara, softverskih inženjera srednjeg do višeg nivoa, naučnika podataka ili analitičara koji su zainteresovani da pređu na inženjerstvo podataka; ili inženjera podataka koji poznaju srž specifičnih tehnologija, ali žele da razviju sveobuhvatniju perspektivu. Naša sekundarna ciljna publika se sastoji od stejkholdera podataka koji rade uz tehničke praktičare – npr. vođa tehničkog tima sa tehničkom pozadinom koji nadgleda tim inženjera podataka, ili direktor skladišta podataka koji želi da migrira sa lokalne tehnologije na rešenje bazirano u oblaku.

Idealno, radoznali ste i želite da učite – zašto biste inače čitali ovu knjigu? Pratite aktuelne tehnologije i trendove podataka čitajući knjige i članke o skladištenju podataka/jezerima podataka, paketnim i strimovanim sistemima, orkestraciji, modelovanju, upravljanju, analizi, razvojima u tehnologijama u oblaku, itd. Ova knjiga će vam pomoći da ispratite ono što ste pročitali u kompletnu sliku inženjerstva podataka preko tehnologija i paradigmi.

## Preduslovi

Pretpostavljamo dobro poznavanje tipova sistema za upravljanje podacima koji se nalaze u korporativnom okruženju. Pored toga, pretpostavljamo da čitaoci imaju izvesno poznavanje SQL-a i Pajton (Python) jezika (ili nekog drugog programskog jezika), kao i iskustva sa klaud (cloud, oblak) uslugama.

Brojni resursi su dostupni budućim inženjerima podataka za vežbanje Pajtona i SQL-a. Na raspolaganju je mnoštvo besplatnih onlajn resursa (blogovi, tutorijanl sajtovi, YouTube videi), a svake godine se objavljuje veliki broj novih Pajton knjiga.

Oblak pruža neviđene mogućnosti za sticanje praktičnog iskustva sa alatima za upravljanje podacima. Predlažemo da budući inženjeri podataka otvore naloge na klaud uslugama poput AWS, Azure, Google Cloud Platform, Snowflake, Databricks, itd. Imajte na umu da mnoge od ovih platformi imaju opcije *besplatnog nivoa*, ali čitaoci bi trebalo da pažljivo prate troškove i rade sa malim količinama podataka i pojedinačnim klaster čvorovima tokom učenja.

Razvijanje poznavanja korporativnih sistema za upravljanje podacima izvan korporativnog okruženja ostaje težak zadatak ito stvara određene prepreke za buduće inženjere podataka koji još uvek nisu dobili svoj prvi posao vezan za podatke. Ova knjiga može da pomogne. Predlažemo da početnici u oblasti podataka prvo pročitaju knjižu na visokom nivou i zatim pregledaju materijale u odeljku *Dodatni izvori* na kraju svakog poglavlja. Tokom drugog čitanja, zabeležite sve nepoznate termine i tehnologije. Možete koristiti Google, Vikipediju, blogove, YouTube videe i sajtove dobavljača da biste se upoznali sa novim terminima i popunili praznine u razumevanju. Predlažemo i rečnik Mikro knjige na stranici *mirkoknjiga.rs*.

## Šta ćete naučiti i kako će unaprediti svoje sposobnosti

Ova knjiga ima za cilj da vam pomogne da izgradite čvrstu osnovu za rešavanje realnih problema inženjerstva podataka.

Do kraja ove knjige, razumećete:

- Kako iinženjerstvo podataka utiče na vašu trenutnu ulogu (naučnik za podatke, softverski inženjer ili vođa tima podataka)
- Kako da presečete marketinšku pompu i odaberete prave tehnologije, arhitekturu podataka i procese
- Kako da koristite životni ciklus inženjerstva podataka za dizajniranje i izgradnju robusne arhitekture
- Najbolje prakse za svaku fazu životnog ciklusa podataka

Bićete u mogućnosti da:

- Ugradite principe inženjerstva podataka u vašu trenutnu ulogu (naučnik za podatke, analitičar, softverski inženjer, vođa tima podataka, itd.)
- Sastavite razne klaud tehnologije kako biste zadovoljili potrebe klijenata nizvodnih podataka
- Procenite probleme inženjerstva podataka sa sveobuhvatnim okvirom najboljih praksi
- Ugradite upravljanje podacima i bezbednost kroz ceo životni ciklus inženjerstva podataka

## Navigacija kroz ovu knjigu

Knjiga se sastoji od četiri dela:

- Deo I, „Osnove i gradivni blokovi“
- Deo II, „Životni ciklus inženjerstva podataka u dubini“
- Deo III, „Bezbednost, privatnost i budućnost inženjerstva podataka“
- Dodaci A i B: koji obrađuju serijalizaciju i kompresiju, odnosno mrežno povezivanje u oblaku

U Delu I, počinjemo definisanjem inženjerstva podataka u Poglavlju 1, zatim mapiramo životni ciklus inženjerstva podataka u Poglavlju 2. U Poglavlju 3, govorimo o *dobroj arhitekturi*. U Poglavlju 4, uvodimo okvir za izbor prave tehnologije – iako često vidimo da se tehnologija i arhitektura poistovećuju, ovo su zapravo veoma različite teme.

Deo II nadograđuje se na Poglavlje 2 da bi detaljno obuhvatio životni ciklus inženjerstva podataka; svaka faza životnog ciklusa – generisanje, skladištenje, unos, transformacija i servisiranje podataka – obrađena je u svom poglavlju.

Deo II se može smatrati srcem knjige, a ostala poglavlja postoje da podrže ključne ideje obrađene ovde.

Deo III pokriva dodatne teme. U Poglavlju 10, govorimo o *bezbednosti i privatnosti*. Iako je bezbednost oduvek bila važan deo profesije inženjerstva podataka, ona je postala još kritičnija sa porastom profiterskog hakovanja i sajber napada sponzorisanih od strane država. Šta možemo reći o privatnosti? Era korporativnog nihilizma privatnosti je završena – nijedna kompanija ne želi da vidi svoje ime u naslovu članka o neadekvatnoj praksi privatnosti. Nesmotren tretman ličnih podataka može imati značajne pravne posledice sa pojavom GDPR-a, CCPA i drugih propisa. Ukratko, bezbednost i privatnost moraju biti apsolutni prioriteti u bilo kom radu inženjerstva podataka.

Tokom rada u inženjerstvu podataka, tokom istraživanja za ovu knjigu i intervjuisanja brojnih stručnjaka, dosta smo razmišljali o tome kuda ovo polje ide u bliskoj i daljoj budućnosti. Poglavlje 11 iznosi naše veoma spekulativne ideje o budućnosti inženjerstva podataka. Po svojoj prirodi, budućnost je klizava stvar. Vreme će pokazati da li su neke od naših ideja tačne. Voleli bismo da čujemo od naših čitalaca kako se njihove vizije budućnosti slažu ili razlikuju od naših.

U dodacima, obrađujemo nekoliko tehničkih tema koje su izuzetno relevantne za svakodnevnu praksu inženjerstva podataka, ali nisu mogle biti uklopljene u glavno telo teksta. Konkretno, inženjeri moraju da razumeju serijalizaciju i kompresiju (Dodatak A) kako bi radili direktno sa datotekama podataka i procenili razmatranja performansi u sistemima za podatke, a mrežno povezivanje na oblaku (Dodatak B) je kritična tema kako se inženjerstvo podataka seli u oblak.

## Konvencije korišćene u ovoj knjizi

U ovoj knjizi koriste se sledeće tipografske konvencije:

*Kurziv*
: Označava nove termine, URL-ove, email adrese, nazive datoteka i ekstenzije datoteka

`Konstantna širina`
: Koristi se za liste programa, kao i unutar pasusa da se odnosi na elemente programa poput imena promenljivih ili funkcija, baza podataka, tipova podataka, promenljivih okruženja, izjava i ključnih reči


Ovaj element označava savet ili sugestiju.


Ovaj element označava opštu napomenu.


Ovaj element ukazuje na upozorenje ili oprez.

## Kako nas kontaktirati

Komentare i pitanja u vezi ove knjige možete uputiti izdavaču na:

MIKRO KNJIGA, DOO
Kneza Višeslava 34
11030 Beograd
Srbija +381 11 7702-883
redakcija@mikroknjiga.rs

Ispravke i dodatne informacije origninala knjige nalaze se na:
*https://oreil.ly/fundamentals-of-data*.

## Zahvalnice

Kada smo počeli da pišemo ovu knjigu, mnogi su nas upozorili da nas očekuje težak zadatak. Knjiga poput ove ima puno pokretnih delova, a zbog svog sveobuhvatnog pogleda na polje inženjerstva podataka, zahtevala je mnogo istraživanja, intervjua, diskusija i dubokog razmišljanja. Nećemo tvrditi da smo uhvatili svaku nijansu inženjerstva podataka, ali se nadamo da će rezultati odjeknuti kod vas. Brojne ličnosti su doprinele našim naporima i zahvalni smo na podršci koju smo dobili od mnogih stručnjaka.

Prvo, hvala našem neverovatnom timu tehničkih recenzenata. Oni su se probili kroz mnoga čitanja i dali neprocenjive (a često i brutalno iskrene) povratne informacije. Ova knjiga bi bila samo deo onoga što jeste bez njihovih napora: Bill Inmon, Andy Petrella, Matt Sharp, Tod Hansmann, Chris Tabb, Danny Lebzyon, Martin Kleppman, Scott Lorimor, Nick Schrock, Lisa Steckman, Veronika Durgin i Alex Woolford.

Drugo, imali smo jedinstvenu priliku da razgovaramo sa vodećim stručnjacima u oblasti podataka na našim uživo emitovanim emisijama, podkastovima, okupljanjima i beskrajnim privatnim pozivima. Njihove ideje su pomogle u oblikovanju naše knjige. Previše je ljudi da bi ih pojedinačno imenovali, ali ovo je ekipa koju želimo da istaknemo: Jordan Tigani, Zhamak Dehghani, Ananth Packkildurai, Shruti Bhat, Eric Tschetter, Benn Stancil, Kevin Hu, Michael Rogove, Ryan Wright, Adi Polak, Shinji Kim, Andreas Kretz, Egor Gryaznov, Chad Sanderson, Julie Price, Matt Turck, Monica Rogati, Mars Lan, Pardhu Gunnam, Brian Suk, Barr Moses, Lior Gavish, Bruno Aziza, Gian Merlino, DeVaris Brown, Todd Beauchene, Tudor Girba, Scott Taylor, Ori Rafael, Lee Edwards, Bryan Offutt, Ollie Hughes, Gilbert Eijkelenboom, Chris Bergh, Fabiana Clemente, Andreas Kretz, Ori Reshef, Nick Singh, Mark Balkenende, Kenten Danas, Brian Olsen, Rhaghu Murthy, Greg Coquillo, David Aponte, Demetrios Brinkmann, Sarah Catanzaro, Michel Tricot, Levi Davis, Ted Walker, Carlos Kemeny, Josh Benamram, Chanin Nantasenamat,

George Firican, Jordan Goldmeir, Minhaaj Rehmam, Luigi Patruno, Vin Vashista, Danny Ma, Jesse Anderson, Alessya Visnjic, Vishal Singh, Dave Langer, Roy Hasson, Todd Odess, Che Sharma, Scott Breitenother, Ben Taylor, Thom Ives, John Thompson, Brent Dykes, Josh Tobin, Mark Kosiba, Tyler Pugliese, Douwe Maan, Martin Traverso, Curtis Kowalski, Bob Davis, Koo Ping Shung, Ed Chenard, Matt Sciorma, Tyler Folkman, Jeff Baird, Tejas Manohar, Paul Singman, Kevin Stumpf, Willem Pineaar i Michael Del Balso iz Tectona, Emma Dahl, Harpreet Sahota, Ken Jee, Scott Taylor, Kate Strachnyi, Kristen Kehrer, Taylor Miller, Abe Gong, Ben Castleton, Ben Rogojan, David Mertz, Emmanuel Raj, Andrew Jones, Avery Smith, Brock Cooper, Jeff Larson, Jon King, Holden Ackerman, Miriah Peterson, Felipe Hoffa, David Gonzalez, Richard Wellman, Susan Walsh, Ravit Jain, Lauren Balik, Mikiko Bazeley, Mark Freeman, Mike Wimmer, Alexey Shchedrin, Mary Clair Thompson, Julie Burroughs, Jason Pedley, Freddy Drennan, Jason Pedley, Kelly i Matt Phillipps, Brian Campbell, Faris Chebib, Dylan Gregerson, Ken Myers, Jake Carter, Seth Paul, Ethan Aaron i mnogi drugi.

Ako niste posebno pomenuti, nemojte to shvatiti lično. Vi znate ko ste. Javite nam se i spomenućemo vas u sledećem izdanju.

Želeli bismo da se zahvalimo timu Ternary Data (Colleen McAuley, Maike Wells, Patrick Dahl, Aaron Hunsaker i drugima), našim studentima i beskrajno mnogim ljudi širom sveta koji su nas podržali. To je sjajan podsetnik da je svet zaista mali.

Rad sa O'Reilly ekipom je bio fantastičan! Posebna zahvalnost ide Jess Haberman na poverenju u nas tokom procesa predlaganja knjige, našim neverovatnim i izuzetno strpljivim urednicama razvoja Nicole Taché i Michele Cronin za dragocene uredničke sugestije, povratne informacije i podršku. Hvala i sjajnom proizvodnom timu u O'Reilly-ju (Greg i ekipa).

Joe bi želeo da se zahvali svojoj porodici – Cassie, Milo i Ethan – što su mu dozvolili da napiše knjigu. Morali su da izdrže mnogo toga, a Joe obećava da nikada više neće napisati knjigu. ;)

Matt bi želeo da se zahvali svojim prijateljima i porodici na njihovom trajnom strpljenju i podršci. On se i dalje nada da će se Seneca udostojiti da da petogodišnju recenziju nakon mnogo napora i propuštenog vremena sa porodicom tokom praznika.