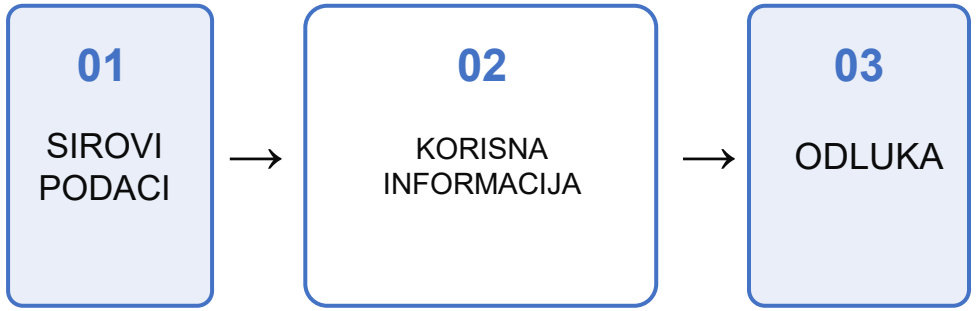
PODACI KAO OSNOVA ANALIZE

Od sirovih činjenica do korisne odluke



Tradicionalni podaci, Big Data, priprema podataka, BI i mašinsko učenje

Predavač: dr Dušan Stefanović

Podaci smanjuju neizvesnost pri odlučivanju

ZAŠTO SU PODACI VAŽNI

Bez podataka

odluke se oslanjaju uglavnom na pretpostavke, iskustvo i intuiciju

Sa relevantnim podacima

1. merimo trenutno stanje
2. proveravamo pretpostavke
3. pratimo posledice odluke

Podaci ne donose odluku umesto čoveka — daju osnovu da odluka bude proverljiva.

Podatak je digitalno zabeležena činjenica

OSNOVNA DEFINICIJA

PODATAK

zapis o objektu, događaju ili merenju

PRIMERI

1. cena proizvoda: 1.299 RSD
2. vreme prijave: 09:42
3. ocena korisnika: 8/10
4. fotografija proizvoda

Značenje podatka nastaje tek kada znamo kontekst, jedinicu i svrhu korišćenja.

Vrednost nastaje transformacijom podataka

OD ČINJENICE DO AKCIJE

SIROVI PODACI
pojedinačni zapisi
bez provere

→

OBRADA
čišćenje, povezivanje
i izračunavanje

→

INFORMACIJA
značenje i obrazac
u kontekstu

→

ODLUKA
izbor akcije
i merenje efekta

Obrada ne dodaje vrednost samo formatiranjem — ona omogućava pouzdano tumačenje.

Podaci se praktično posmatraju u dva okruženja

TRADICIONALNI PODACI I BIG DATA

TRADICIONALNI PODACI

pretežno strukturirani
organizovani u tabelama
obrađuju se klasičnim bazama i alatima

VELIKI PODACI — BIG DATA

veliki obim, raznovrsni formati
brzo nastajanje i obrada
često distribuirano skladištenje

Granica nije samo u broju gigabajta; bitno je da li postojeći alati mogu efikasno da obrade podatke.

Tradicionalni podaci imaju jasnu i stabilnu strukturu

TRADICIONALNI PODACI

Tipičan oblik

ID	KUPAC	GRAD	KUPOVINE
101	Ana	Niš	4
102	Marko	Beograd	7
103	Sara	Novi Sad	2

Karakteristike

1. redovi i kolone
2. poznati tipovi polja
3. relacione baze podataka
4. SQL upiti i izveštaji

Tabela je prirodan prikaz kada su atributi unapred poznati i podaci imaju stabilnu šemu.

Big Data zahteva skaliranje obrade i skladištenja

VELIKI PODACI

Više izvora

- aplikacije
- senzori
- slike i video
- logovi
- transakcije

→

DISTRIBUIRANI SISTEM

podaci se dele između više računara

obrada se izvršava paralelno

→

REZULTAT

brži uvid
obrada u realnom vremenu
skaliranje kapaciteta

Big Data je problem arhitekture i brzine obrade, a ne samo veliki fajl.

Tri V objašnjavaju zašto su veliki podaci zahtevni

VOLUME — VARIETY — VELOCITY

VOLUME

Koliko podataka postoji?
TB, PB, EB

VARIETY

U kojim formatima dolaze?
tabele, tekst, slika, zvuk

VELOCITY

Koliko brzo nastaju i moraju da se obrade?
serijski tok, realno vreme

Skup podataka može biti izazovan zbog jednog V ili njihove kombinacije.

Tradicionalni i veliki podaci traže različito skaliranje

POREĐENJE

Osobina	Tradicionalni podaci	Big Data
Struktura	pretežno strukturirana	strukturirana, polustrukturirana i nestrukturirana
Obim	upravljiv na jednom sistemu	često zahteva distribuirano skladištenje
Brzina	periodična obrada	često kontinuirana ili u realnom vremenu
Tipični alati	relacione baze, SQL	distribuirani sistemi, tokovi podataka
Primer	tabela kupaca	klikovi, video, senzori i logovi zajedno

Izbor tehnologije zavisi od strukture, obima, brzine i potrebnog vremena odgovora.

Big Data nastaje kada digitalni događaji stalno proizvode zapise

PRIMERI IZ PRAKSE

DRUŠTVENA PLATFORMA

profili i veze
fotografije i video
poruke i reakcije
klikovi i logovi

Velika raznovrsnost i brzina nastanka

FINANSIJSKO TRGOVANJE

cena i količina
ponude i tražnje
transakcije u kratkim intervalima

Veliki obim i potreba za veoma brzom obradom

Isti 3V okvir pomaže da razumemo potpuno različite industrije.

Poslovno pitanje određuje šta radimo sa podacima

MESTO PODATAKA U PROCESU

PODACI	→	BI	→	MODEL	→	ODLUKA
relevantni i organizovani zapisi		šta se dogodilo i gde?		šta se može očekivati?		koju akciju biramo?

Bez relevantnih podataka ni dashboard ni model ne mogu dati pouzdan rezultat.

Vremenska osa razdvaja objašnjenje i predviđanje

PROŠLOST — SADAŠNJOST — BUDUĆNOST



Analiza prošlosti stvara osnovu, ali predviđanje zahteva model i proveru njegovih rezultata.

Prikupljanje podataka je prvi operativni korak

DATA COLLECTION

AKTIVNO PRIKUPLJANJE

ankete i upitnici
intervjui
merenja i eksperimenti
ručni unos

AUTOMATSKO PRIKUPLJANJE

transakcioni sistemi
logovi aplikacija
senzori
kolačići i događaji na sajtu

Način prikupljanja utiče na pristrasnost, tačnost, privatnost i kasniju analizu.

Preprocessing priprema podatke za pouzdanu analizu

DATA PRE-PROCESSING

PRIKUPLJANJE
sirovi zapisi
iz različitih izvora

→

PROVERA
format, opseg, tip
i potpunost

→

ČIŠĆENJE
greške, duplikati
i nedostajuće vrednosti

→

PRIPREMA
spajanje,
transformacija
i format za analizu

Model ne popravlja loše podatke — najčešće samo brže proizvodi nepouzdan rezultat.

Problemi kvaliteta mogu potpuno promeniti zaključak

ŠTA PROVERAVAMO

NEMOGUĆA VREDNOST	starost = 932	POGREŠNO POLJE	ime = Ujedinjeno Kraljevstvo
NEDOSTAJE PODATAK	starost nije uneta	NEUJEDNAČEN ZAPIS	Niš / Nis / NIŠ
DUPLIKAT	isti kupac unet dva puta	PRISTRASAN UZORAK	80% odgovora iz jedne grupe

Greška u podatku postaje greška u izveštaju, analizi i modelu ako se ne otkrije na vreme.

Tip promenljive određuje dozvoljene operacije

CLASS LABELING I TIPOVI PODATAKA

NUMERIČKE PROMENLJIVE

izražavaju količinu ili merenje

možemo računati zbir, prosek, razliku i odnos

primer: prihod, temperatura, broj kupovina

KATEGORIJALNE PROMENLJIVE

označavaju grupu, vrstu ili identitet

računamo učestalost, ali prosek često nema smisla

primer: zanimanje, grad, tip uređaja

Izgled vrednosti nije dovoljan; tip određujemo prema značenju i analitičkoj ulozi.

Numeričke i kategorijalne vrednosti odgovaraju na različita pitanja

KAKO IH KORISTIMO

Pitanje	Numerička promenljiva	Kategorijalna promenljiva
Šta predstavlja?	količinu ili merenje	grupu, klasu ili oznaku
Tipična analiza	prosek, zbir, medijana, rasipanje	broj i udeo po kategoriji
Primer	broj kupovina = 7	grad = Niš
Dobar upit	kolika je prosečna kupovina?	koji grad ima najviše kupaca?

Pogrešno određen tip promenljive vodi pogrešnoj obradi i besmislenim statistikama.

ID i datum pokazuju zašto značenje ima prednost nad formatom

GRANIČNI PRIMERI

ID KUPCA = 103

zapisan je brojem

ali predstavlja identitet

prosečan ID nema korisno značenje

BROJ ŽALBI = 3

zapisan je brojem

predstavlja količinu

zbir i prosek imaju značenje

DATUM = 01.09.2026.

predstavlja vreme

možemo izdvojiti mesec, dan ili razliku

nije automatski kategorijalna promenljiva

Format zapisa, statistički tip i poslovno značenje nisu ista stvar.

Čišćenje podataka uklanja ili označava nedoslednosti

DATA CLEANSING / DATA CLEANING

OTKRIJ	→	PROVERI	→	ISPRAVI	→	DOKUMENTUJ
pravilo, profilisanje ili pregled		uporedi sa izvorom i poslovnim pravilom		standardizuj, spoji ili označi kao nevažeće		zapiši promenu i razlog

Dobra korekcija je ponovljiva, proverljiva i ne skriva izvorni problem.

Nedostajuća vrednost se ne rešava jednim univerzalnim pravilom

MISSING VALUES

Nedostaje starost kupca

1. RAZUMI	→	2. PROCENI	→	3. IZABERI	→	4. PROVERI
zašto podatak nedostaje?		koliko zapisa je pogođeno?		dopuna, posebna oznaka ili isključenje		da li je zaključak promenjen?

Zamena prosekom može biti praktična, ali može smanjiti varijabilnost i uvesti pristrasnost.

Balansiranje sprečava da većinska grupa potisne manjinsku

BALANCING

Pre balansiranja



80% grupa A | 20% grupa B

Posle izabranog postupka



50% grupa A | 50% grupa B

Mogući postupci: pažljivo uzorkovanje, ponderisanje ili promene u obuci modela

Cilj nije uvek odnos 50:50, već pouzdano predstavljanje problema koji rešavamo.

Mešanje redosleda uklanja slučajne obrasce prikupljanja

DATA SHUFFLING

Pre mešanja



Posle mešanja



Redosled više ne zavisi od vremena ili grupe u kojoj su zapisi uneseni.

Shuffling menja redosled, ali ne rešava duplikate, pogrešne vrednosti ili pristrasan izbor uzorka.

Maskiranje štiti osetljive podatke tokom korišćenja

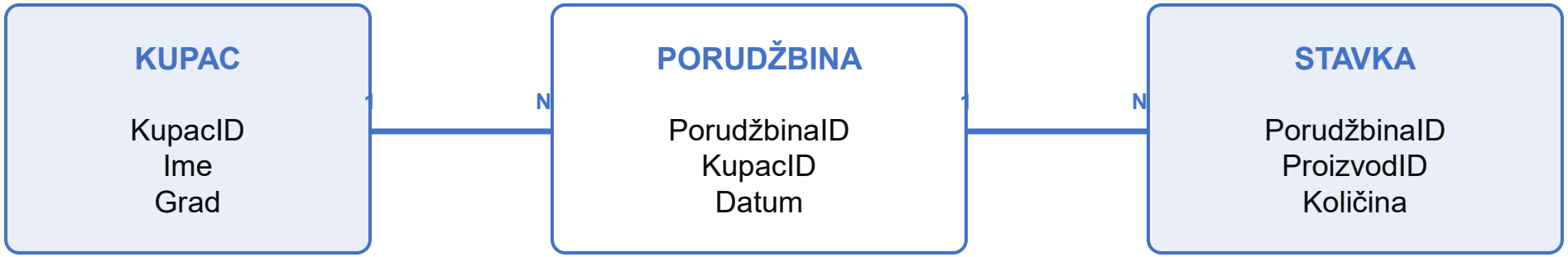
DATA MASKING

IZVORNI PODATAK		PRAVILO MASKIRANJA		MASKIRANI PODATAK
JMBG: 0101990... e-mail: ana@firma.rs	→	zamena, skraćivanje ili pseudonimizacija	→	JMBG: 0101****** e-mail: a***@firma.rs

Maskiranje smanjuje izloženost, ali ne zamenjuje kontrolu pristupa, pravni osnov i bezbednost sistema.

ER dijagram i relaciona šema prikazuju strukturu baze

TRADICIONALNA ARHITEKTURA PODATAKA

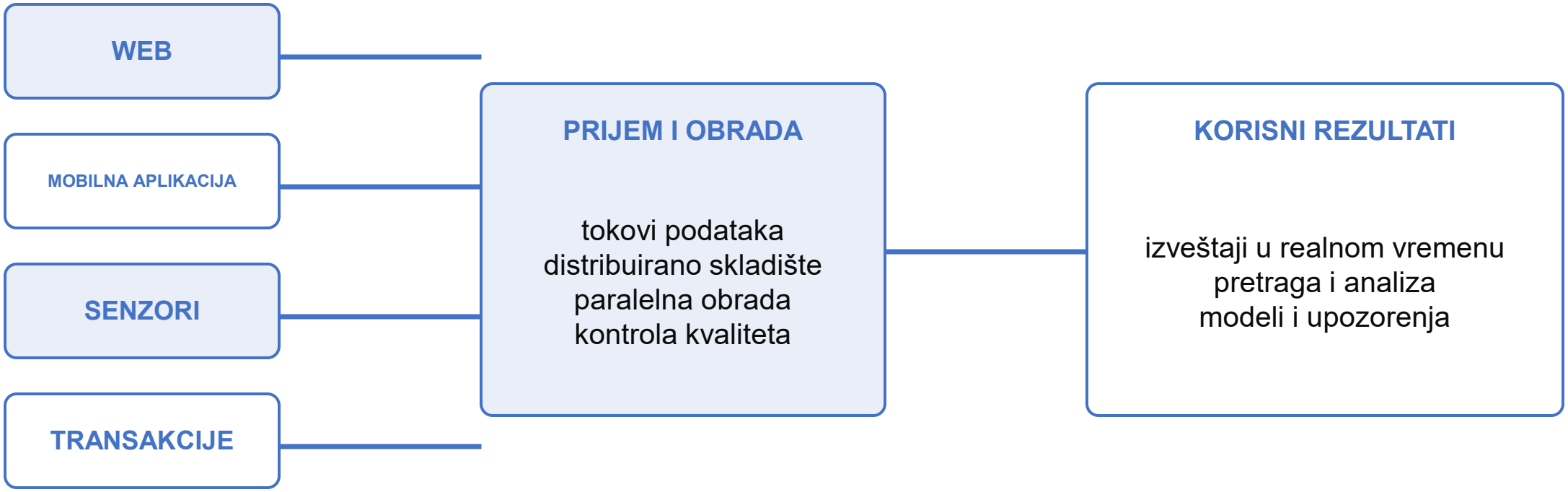


ER dijagram opisuje entitete i veze. Relaciona šema pokazuje tabele, ključeve i način povezivanja.

Struktura baze omogućava da se podaci povežu bez nepotrebnog ponavljanja.

Big Data povezuje mnogo izvora kroz tok obrade

DISTRIBUIRANA ARHITEKTURA



Složenost velikih podataka dolazi iz koordinacije izvora, formata, brzine i infrastrukture.

Primer prodavnice povezuje sve korake u jedan proces

PRAKTIČAN SCENARIO

PRIKUPLJANJE	→	PRIPREMA	→	BI I ANALIZA	→	MODEL I ODLUKA
kupovine, posete, žalbe i ocene zadovoljstva		spajanje kupaca čišćenje i nedostajuće vrednosti		gde pada prodaja i koji segment ima problem		rizik odlaska i ciljana ponuda

Efekat odluke ponovo postaje podatak i pokreće novi ciklus analize.

Najveća vrednost nastaje kada su podaci, analiza, model i poslovna akcija povezani.

Kvalitetan rezultat počinje mnogo pre algoritma

ZAVRŠNI PREGLED

PODACI

relevantni, pravilno prikupljeni i opisani

PRIPREMA

provera, čišćenje, tipovi, balans i zaštita

ANALIZA

BI za pregled; statistika i ML za modele

ODLUKA

rezultat se primenjuje, proverava i meri

Loš ulaz se ne pretvara u dobar zaključak samo izborom naprednijeg algoritma.

Pitanje 1

PROVERA ZNANJA

Koja tvrdnja najbolje opisuje podatak?

A. To je svaka odluka koju menadžer donese

B. To je digitalno zabeležena činjenica kojoj je za tumačenje potreban kontekst

C. To je isključivo broj smešten u tabeli

D. To je rezultat koji je uvek tačan

Izaberite jedan odgovor i obrazložite izbor.

Pitanje 2

PROVERA ZNANJA

Koja tri svojstva čine osnovni 3V model velikih podataka?

A Volume, Variety i Velocity

B Value, Vision i Verification

C Volume, Validity i Visualization

D Velocity, Version i Value

Izaberite jedan odgovor i obrazložite izbor.

Pitanje 3

PROVERA ZNANJA

Zašto ID kupca, iako je zapisan brojem, nije tipična numerička promenljiva?

A. Zato što ID ne sme da se čuva u bazi

B. Zato što prosečna ili zbirna vrednost ID-ja nema poslovno značenje

C. Zato što svi brojevi predstavljaju kategorije

D. Zato što ID uvek mora biti tekst dužine 30

Izaberite jedan odgovor i obrazložite izbor.

Pitanje 4

PROVERA ZNANJA

Koji postupak nasumično menja redosled posmatranja u skupu podataka?

A Data masking

B Data cleansing

C Data shuffling

D Data labeling

Izaberite jedan odgovor i obrazložite izbor.

Pitanje 5

PROVERA ZNANJA

Koje pravilo je najispravnije pri izboru statističkog metoda ili ML-a?

A. ML se koristi samo kada postoje veliki podaci

B. Statistički modeli se koriste samo na malim tabelama

C. Uvek treba izabrati najsloženiji algoritam

D. Izbor zavisi od problema, podataka, tačnosti, objašnjivosti i troška

Izaberite jedan odgovor i obrazložite izbor.