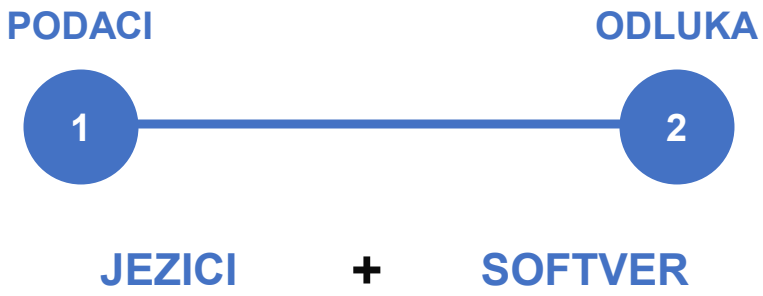
ALATI ZA OBRADU PODATAKA

Programski jezici i softverski alati



Kako izabrati pravi alat i povezati ga sa poslovnim problemom

Predavač: dr Dušan Stefanović

Alat pretvara ponovljiv posao u proces

ZAŠTO SU ALATI VAŽNI

1 RUČNI POSAO

isti koraci se ponavljaju

→

2 ALAT

koraci se definišu jednom

→

3 REZULTAT

analiza se ponavlja pouzdano

Najveći značaj alata je ponovljivost, brzina i smanjenje ručnih grešaka.

Jezici daju slobodu, softver ubrzava rad

DVE GRUPE ALATA

PROGRAMSKI JEZICI

Python • R • SQL
MATLAB • Java • Scala
C/C++ • JavaScript

prilagođavanje • automatizacija • integracija

SOFTVERSKI ALATI

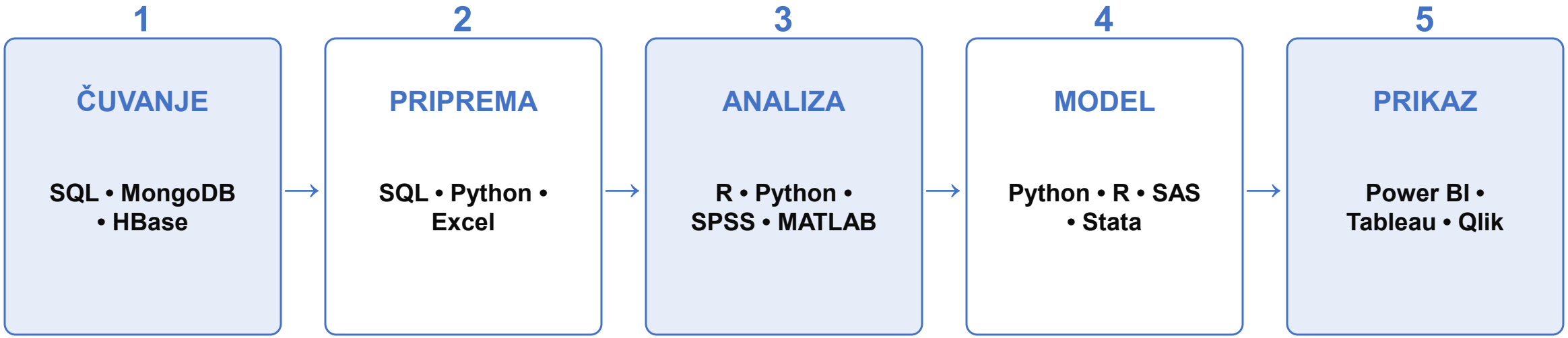
Excel • SPSS • Stata • EViews
Power BI • Tableau • Qlik • SAS
Hadoop • HBase • MongoDB

brzo usvajanje • gotove funkcije • standardizovan rad

U praksi se ove dve grupe najčešće kombinuju.

Svaka faza obrade ima svoje alate

TOK OBRADE PODATAKA



Jedan projekat je lanac alata, a ne izbor samo jednog programa.

Pravi alat zavisi od šest pitanja

KRITERIJUMI IZBORA

1. Koliki je obim podataka?
2. Da li obrada mora biti automatska?
3. Da li je rezultat izveštaj ili model?
4. Koliko brzo rezultat mora biti dostupan?
5. Koja znanja poseduje tim?
6. Koliki su trošak i zahtevi infrastrukture?

Najpopularniji alat nije automatski i najpogodniji alat.

Python povezuje gotovo ceo analitički proces

PROGRAMSKI JEZICI

PYTHON

obrada, automatizacija i ML

ZAŠTO JE VAŽAN

Jedan jezik može da učita podatke, očisti ih, napravi model i poveže rezultat sa aplikacijom.

JEDNOSTAVAN PRIMER

Prodavnica svakog jutra automatski izračunava rizik odlaska kupaca.

OGRANIČENJE

Potrebno je programersko znanje i pažljivo upravljanje bibliotekama i okruženjem.

Značaj Pythona je širina primene i lako povezivanje različitih sistema.

Python automatizuje ponovljivu obradu

PRIMER: DNEVNA PRODAJA

PYTHON SKRIPTA

```
1  učitaj nove račune
2  ukloni duplikate
3  izračunaj prodaju po
proizvodu
4  sačuvaj rezultat
5  pošalji podatke
kontrolnoj tabli
```

→

REZULTAT

svakog dana isti postupak
bez ručnog kopiranja
uz zapis eventualnih grešaka

Kod postaje vredan kada se pouzdano izvršava nad novim podacima.

R daje snažno okruženje za statističko zaključivanje

PROGRAMSKI JEZICI

R

statistika, istraživanje i grafika

ZAŠTO JE VAŽAN

Veliki broj statističkih metoda dostupan je kroz jezik i specijalizovane pakete.

JEDNOSTAVAN PRIMER

Istraživač proverava da li zadovoljstvo kupaca statistički utiče na lojalnost.

OGRANIČENJE

Za razvoj šire poslovne aplikacije često je potrebno povezivanje sa drugim tehnologijama.

Značaj R-a je brz prelazak od podataka do statističkog modela i grafikona.

Python i R se razlikuju po težištu, ne po vrednosti

POREĐENJE JEZIKA

Pitanje	Python	R
Glavni fokus	širi razvoj sistema	statistika i istraživanje
Automatizacija	veoma pogodna	moguća i efikasna
Statistički rad	bogat ekosistem	prirodno težište jezika
Tipičan izbor	model u aplikaciji	analiza i istraživački izveštaj

Izbor zavisi od projekta

SQL pronalazi tačne podatke u relacionoj bazi

PROGRAMSKI JEZICI

SQL

upiti nad relacionim podacima

ZAŠTO JE VAŽAN

Omogućava filtriranje, spajanje i agregiranje podataka tamo gde su oni sačuvani.

JEDNOSTAVAN PRIMER

BI analitičar spaja kupce i račune i računa prodaju po regionu.

OGRANIČENJE

SQL nije naziv baze; konkretni sistemi su PostgreSQL, SQL Server, MySQL i Oracle.

Značaj SQL-a je da analitičaru daje kontrolisan pristup istorijskim podacima.

Jedan SQL upit pretvara račune u poslovni pokazatelj

PRIMER SQL UPITA

```
SELECT customer_id,
       SUM(amount) AS total_spent
FROM sales
GROUP BY customer_id;
```

→

REZULTAT

jedan red po kupcu
+
ukupna potrošnja

SQL sažima veliki broj transakcija pre dalje analize.

MATLAB je važan kada problem počinje matematičkim modelom

NUMERIČKI ALATI

MATLAB

matrice, simulacije i inženjerski proračuni

ZAŠTO JE VAŽAN

Omogućava brzo eksperimentisanje sa matematičkim formulama, signalima i simulacijama.

JEDNOSTAVAN PRIMER

Inženjer filtrira signal senzora i poredi rezultat sa očekivanim modelom.

OGRANIČENJE

Komercijalna licenca može biti ograničenje; u klasičnim data platformama češće se koristi Python.

Značaj MATLAB-a je bliska veza između matematike, algoritma i simulacije.

Java i Scala grade pouzdane sisteme za obradu velikih podataka

JEZICI ZA VELIKE SISTEME

JAVA

veliki poslovni sistemi
dugotrajni servisi
integracija sa platformama

SCALA

obrada podataka na JVM-u
Apache Spark ekosistem
funkcionalni stil obrade

Njihov značaj je stabilno izvršavanje distribuiranih poslova, ne samo statistika.

C/C++ ubrzavaju jezgro, JavaScript približava model korisniku

SPECIFIČNE ULOGE JEZIKA

Jezik	Poseban značaj	Primer
C/C++	brzina i kontrola resursa	biblioteka za numerički proračun
JavaScript	izvršavanje u veb-okruženju	model radi u pregledaču
Python + C++	jednostavan kod iznad brzog jezgra	ML biblioteka sa optimizovanim operacijama

Brzina jezika je važna, ali kvalitet podataka i algoritma ostaje presudan.

Excel je najbrži put do prve poslovne analize

SOFTVERSKI ALATI

EXCEL

tabele, formule i brzi izveštaji

ZAŠTO JE VAŽAN

Poslovni korisnik može brzo proveriti podatke bez razvoja posebne aplikacije.

JEDNOSTAVAN PRIMER

Menadžer pravi pivot tabelu prodaje po mesecu i regionu.

OGRANIČENJE

Ručni koraci, velike tabele i složene formule otežavaju kontrolu i ponovljivost.

Značaj Excela je dostupnost, ali granice nastaju sa obimom i složenošću.

SPSS pojednostavljuje standardnu statističku analizu

STATISTIČKI SOFTVER

SPSS

ankete i primenjena statistika

ZAŠTO JE VAŽAN

Korisnik kroz gotove procedure brzo sprovodi testove i dobija standardizovane rezultate.

JEDNOSTAVAN PRIMER

Analiza studentske ankete: frekvencije, proseci, pouzdanost skale i regresija.

Značaj SPSS-a je dostupnost statistike korisnicima koji ne razvijaju ceo sistem.

Stata i EViews rešavaju različite istraživačke zadatke

STATISTIKA I EKONOMETRIJA

STATA

statističko i ekonometrijsko istraživanje
panel podaci • regresija • akademski rad

Primer: uticaj obrazovanja na zaradu

EVIEWS

ekonometrijske vremenske serije
trend • sezonalnost • prognoziranje

Primer: prognoza prodaje ili inflacije

Stata je šira istraživačka platforma; EViews je posebno prepoznatljiv po vremenskim serijama.

Power BI povezuje podatke sa svakodnevnom odlukom

POSLOVNA INTELIGENCIJA

POWER BI

modeli podataka i kontrolne table

ZAŠTO JE VAŽAN

Povezuje više izvora, definiše poslovne mere i omogućava redovno praćenje pokazatelja.

JEDNOSTAVAN PRIMER

Direktor prati prihod, maržu, plan i zalihe po regionu.

OGRANIČENJE

Lep grafikon nije dovoljan ako model podataka i poslovne mere nisu ispravni.

Značaj Power BI-ja je distribucija zajedničke verzije poslovnih pokazatelja.

Tableau i Qlik ubrzavaju vizuelno istraživanje

BI PLATFORME

TABLEAU

vizuelna analitika
brzo postavljanje novih pitanja
interaktivne priče iz podataka

QLIK

asocijativno istraživanje
integracija i analitika
poređenje povezanih izbora

Njihov značaj je da korisnik istražuje podatke bez pisanja svakog upita ručno.

SAS standardizuje analitiku u kontrolisanim okruženjima

ANALITIČKE PLATFORME

SAS

statistika, podaci i poslovna analitika

ZAŠTO JE VAŽAN

Objedinjuje obradu, statistiku, izveštavanje i upravljanje analitičkim procesom.

JEDNOSTAVAN PRIMER

Banka primenjuje standardizovanu proceduru za regulatorni izveštaj.

OGRANIČENJE

Komercijalna platforma i specifično okruženje mogu povećati trošak i zavisnost od dobavljača.

Značaj SAS-a je kontrolisan i ponovljiv analitički rad u velikim organizacijama.

Hadoop deli veliki problem na više računara

BIG DATA INFRASTRUKTURA

VELIKI SKUP

podaci ne staju ili se sporo obrađuju na jednom računaru

→

DISTRIBUCIJA

podaci i poslovi dele se između računara

→

REZULTAT

delovi se obrađuju paralelno i objedinjuju

Hadoop nije klasična desktop aplikacija, već okvir za distribuirano čuvanje i obradu.

HBase omogućava brz pristup velikim distribuiranim tabelama

BIG DATA SKLADIŠTENJE

HBASE

distribuirana wide-column baza

ZAŠTO JE VAŽAN

Omogućava čuvanje veoma velikih tabela i pristup pojedinačnim redovima u distribuiranom okruženju.

JEDNOSTAVAN PRIMER

Telekom čuva veliki broj događaja uređaja i pristupa zapisima prema ključu korisnika.

OGRANIČENJE

Nije zamena za svaku relacionu bazu; model podataka mora odgovarati obrascu pristupa.

Značaj HBase-a je skalabilan pristup velikim tabelama, a ne klasično BI izveštavanje.

MongoDB čuva promenljive strukture kao dokumente

DOKUMENTNE BAZE

MONGODB

JSON-slični BSON dokumenti

ZAŠTO JE VAŽAN

Podaci iz aplikacije mogu se čuvati kao ugnježdeni dokumenti bez striktne tabele za svaki detalj.

JEDNOSTAVAN PRIMER

Profil proizvoda sadrži različite osobine, niz slika i ugnježdene podatke o proizvođaču.

OGRANIČENJE

Fleksibilna šema ne znači da dizajn podataka i kontrola kvaliteta nisu potrebni.

Značaj MongoDB-a je fleksibilno modelovanje aplikacionih podataka i horizontalno skaliranje.

Hadoop, HBase i MongoDB ne rešavaju isti problem

POREĐENJE BIG DATA ALATA

Tehnologija	Glavna uloga	Tipičan slučaj
Hadoop	distribuirani okvir i skladištenje	veliki serijski poslovi
HBase	wide-column distribuirana baza	brz pristup redovima velike tabele
MongoDB	dokumentna operativna baza	fleksibilni podaci aplikacije

Slična veličina podataka ne znači da je potreban isti tip sistema.

Internet prodavnica koristi više alata kao jedan sistem

POVEZANI PRAKTIČNI PRIMER

SQL

izdvaja kupovine i kupce

→

PYTHON

čisti podatke i pravi model

→

POWER BI

prikazuje rizik i rezultat akcija

→

MENADŽMENT

bira ponudu i meri efekat

Vrednost nastaje tek kada se tehnički rezultat pretvori u merljivu akciju.

Alati razmenjuju podatke kroz jasno definisane izlaze

ARHITEKTURA TOKA



Svaki izlaz mora imati poznatu strukturu, vreme osvežavanja i odgovornu osobu.

Integracija alata je jednako važna kao izbor pojedinačnog alata.

Izaberite najmanji alat koji pouzdano rešava problem

MATRICA IZBORA

Potreba	Prvi izbor	Zašto
Podaci u relacionoj bazi	SQL	obrada blizu izvora
Brza manja analiza	Excel	malo početno opterećenje
Automatizacija i model	Python	fleksibilnost i integracija
Statističko istraživanje	R / SPSS / Stata	gotove statističke metode
Kontrolna tabla	Power BI / Tableau / Qlik	interaktivno predstavljanje

Složeniji alat ima smisla samo ako rešava stvarno ograničenje.

Sedam poruka

PREGLED LEKCIJE

1. SQL pristupa strukturisanim istorijskim podacima.
2. Python i R obrađuju, analiziraju i modeluju.
3. MATLAB povezuje matematiku i inženjersku simulaciju.
4. Excel i statistički paketi ubrzavaju standardne analize.
5. BI alati pretvaraju rezultate u interaktivne uvide.
6. Big data sistemi dele čuvanje i obradu između računara.
7. **Najbolje rešenje povezuje alate u kontrolisan proces.**

Teorija određuje šta računamo; alat određuje kako to pouzdano izvršavamo.

Pitanje 1

PROVERA ZNANJA

Koja je osnovna uloga SQL-a u analitičkom projektu?

A Izrada prezentacije

B Izdvajanje i agregiranje podataka iz relacione baze

C Obrada slike

D Upravljanje Hadoop klasterom

Izaberite jedan odgovor i obrazložite izbor.

Pitanje 2

PROVERA ZNANJA

Kada je programski jezik važniji od gotovog softvera?

A. Kada je potrebno specifično rešenje i automatizacija

B. Kada ne želimo ponoviti analizu

C. Kada nema podataka

D. Kada je dovoljan jedan ručni grafikon

Izaberite jedan odgovor i obrazložite izbor.

Pitanje 3

PROVERA ZNANJA

Koji alat je posebno usmeren na ekonometrijske vremenske serije?

- **A** Power BI
- **B** MongoDB
- **C** Excel
- **D** EViews

Izaberite jedan odgovor i obrazložite izbor.

Pitanje 4

PROVERA ZNANJA

Šta Hadoop prvenstveno omogućava?

A Distribuirano čuvanje i obradu velikih podataka

B Ručno pravljenje tabela

C Dizajn veb-stranice

D Samo statističke grafikone

Izaberite jedan odgovor i obrazložite izbor.

Praktičan slučaj: osmislite analitički sistem

SAMOSTALNA PRIMENA

SCENARIO

Lanac prodavnica želi dnevno da prati prodaju, otkrije proizvode sa rizikom nestašice i prognozira potražnju za naredni mesec.

1. Odredite gde se podaci čuvaju.
2. Izaberite alat za izdvajanje i čišćenje.
3. Izaberite alat za prognoziranje.
4. Izaberite način prikaza rezultata.
5. Objasnite kako se komponente povezuju.

Dobar odgovor povezuje poslovni cilj, podatke, alat, rezultat i odgovornost.