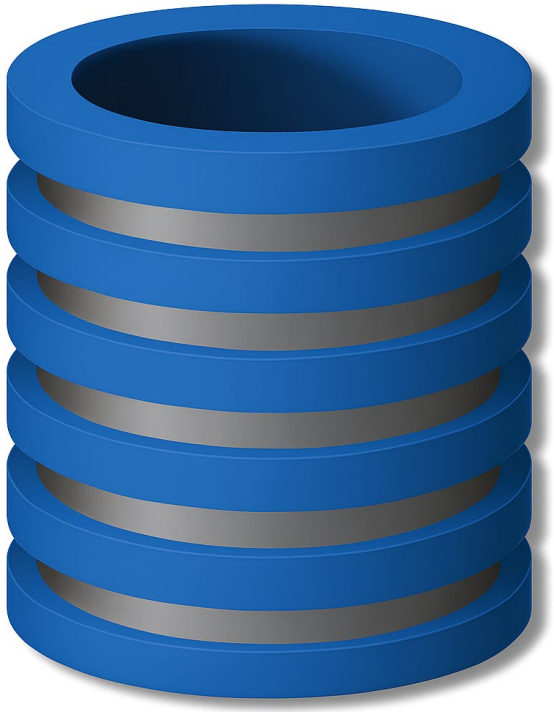


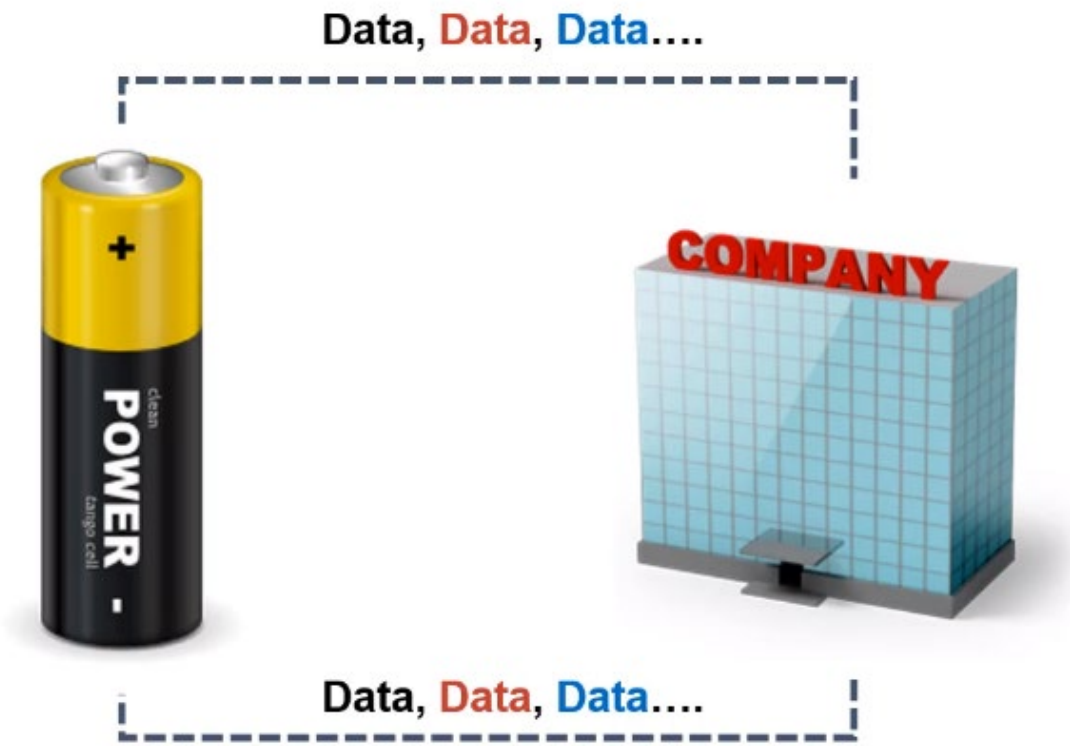
Terminologija u bazama podataka

Predmet: Administriranje Baze Podataka

Predavač: dr Dušan Stefanović

PODACI SU SVUDA OKO NAS

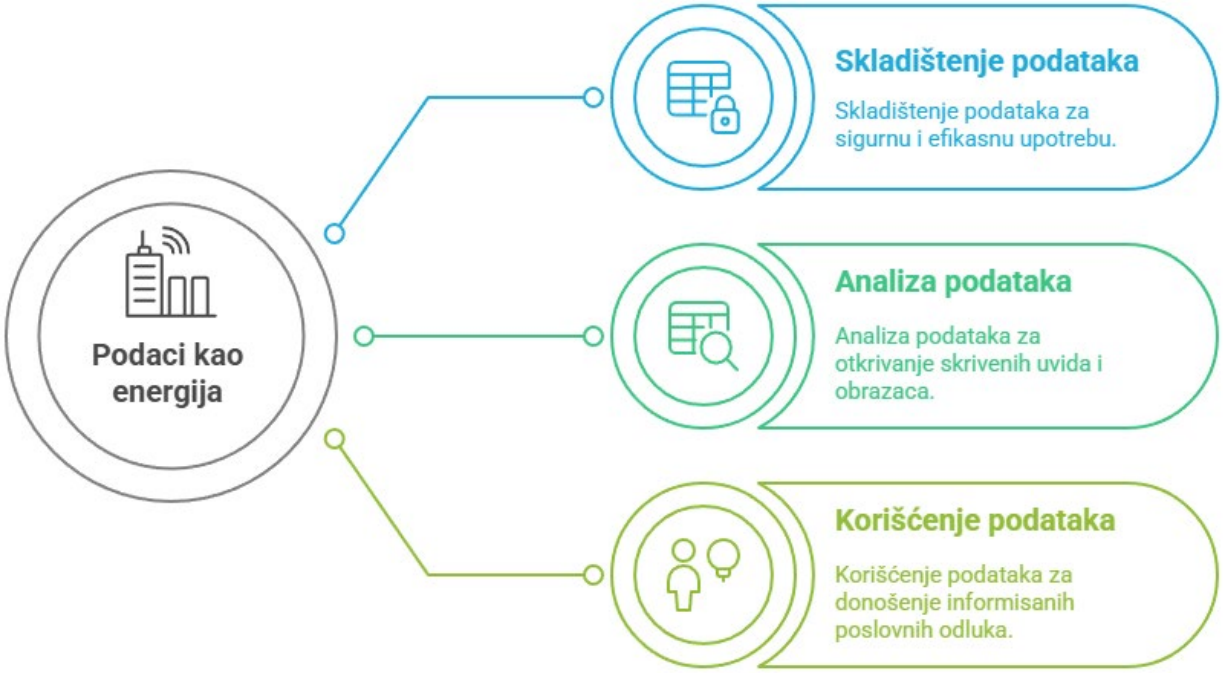
- Podaci su energija koja pokreće kompanije da posluju
- Ako su podaci energija, potrebno je za njih obezbediti mesto gde će se čuvati



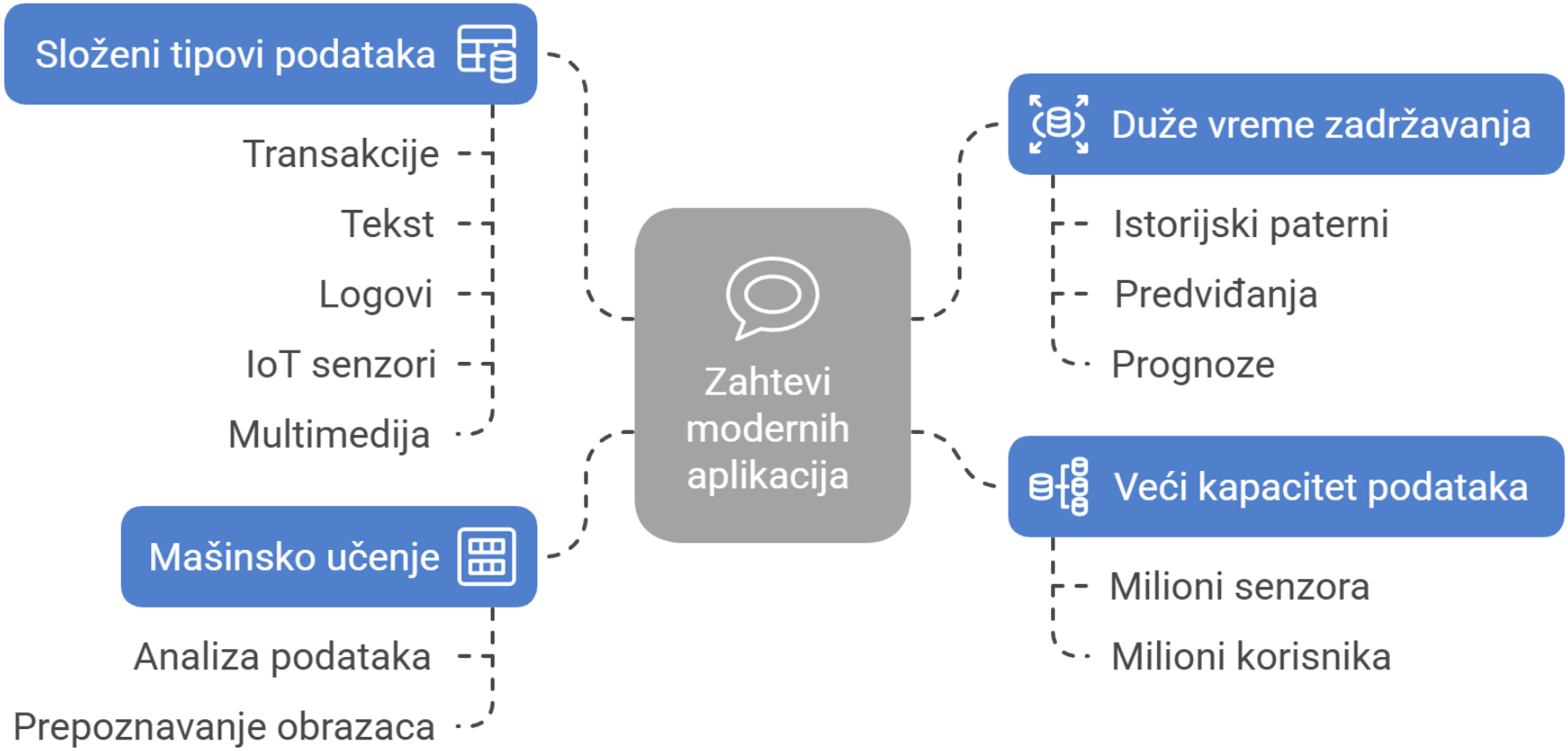
PODACI SU SVUDA OKO NAS

Kontejner za podatke & Engine

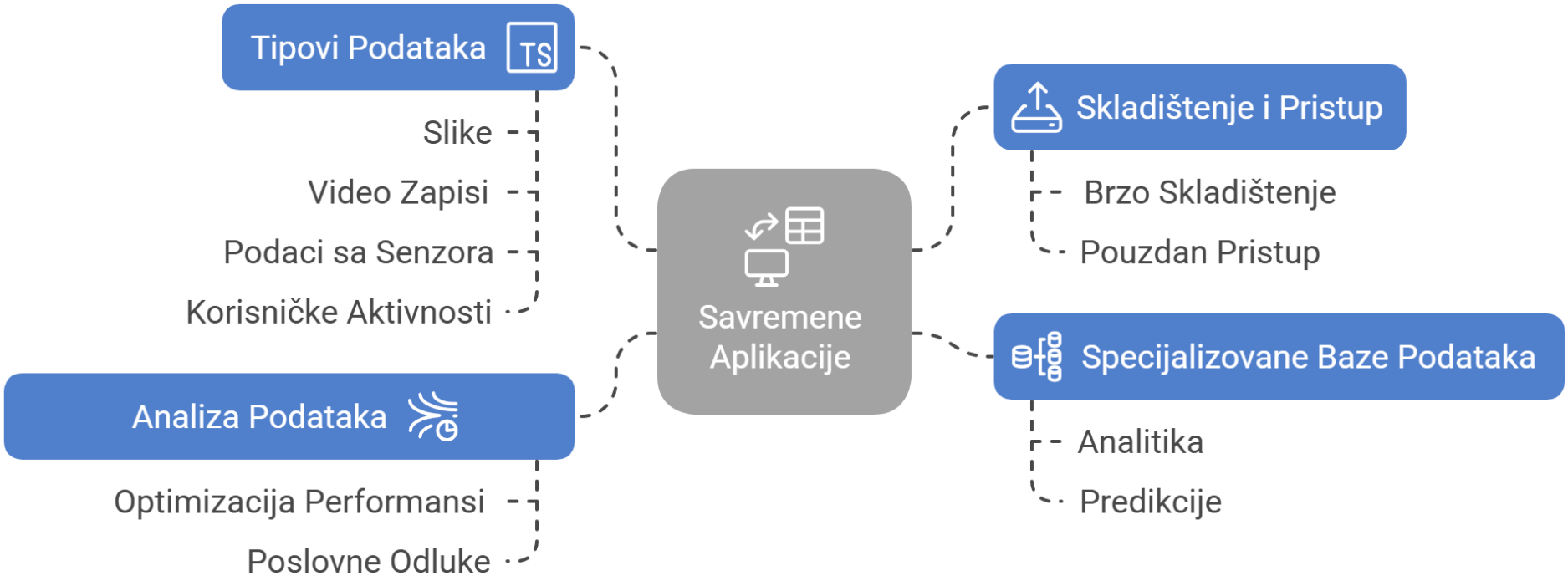
- Fleksibilan
- Brz
- Pouzdan
- Lako upravljiv
- Isplativ



ČUVANJE I OBRADA PODATAKA



ZAHTEVI SAVREMENIH APLIKACIJA



PODACI VELIKIH BROJEVA

800 Terabytes – 2000.

160 Exabytes – 2006. (1EB = 10^{18}B)

4.5 Zettabytes – 2013. (1ZB = 10^{21}B)

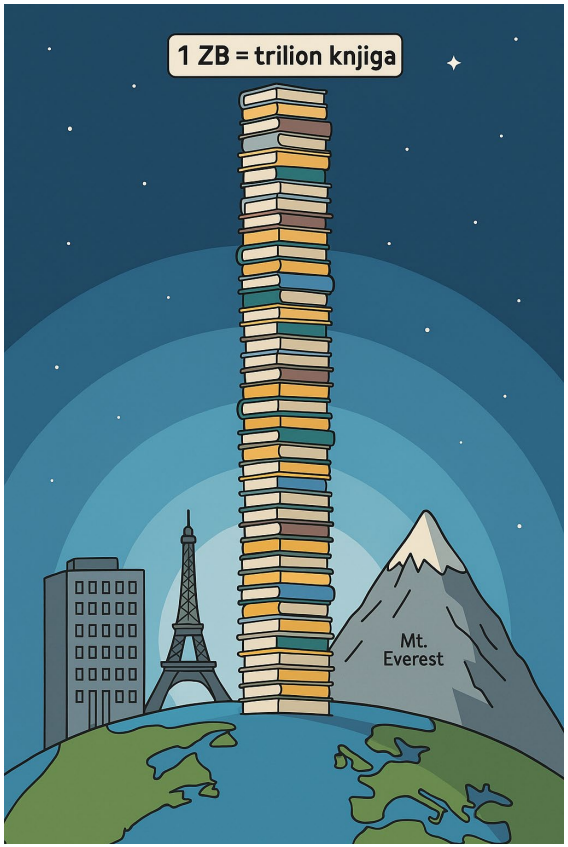
44 Zettabytes – 2020.

163 Zettabytes do 2025.

53 Important Statistics About How Much Data Is Created Every Day in 2024 - Financesonline.com

KOLIKO SU PODACI VELIKI

1 Zettabyte = naslagane knjige visine 5 puta rastojanje od Zemlje do Plutona



KOLIKO SU PODACI VELIKI

8 milliona godina video sadržaja u UHD 8K formatu

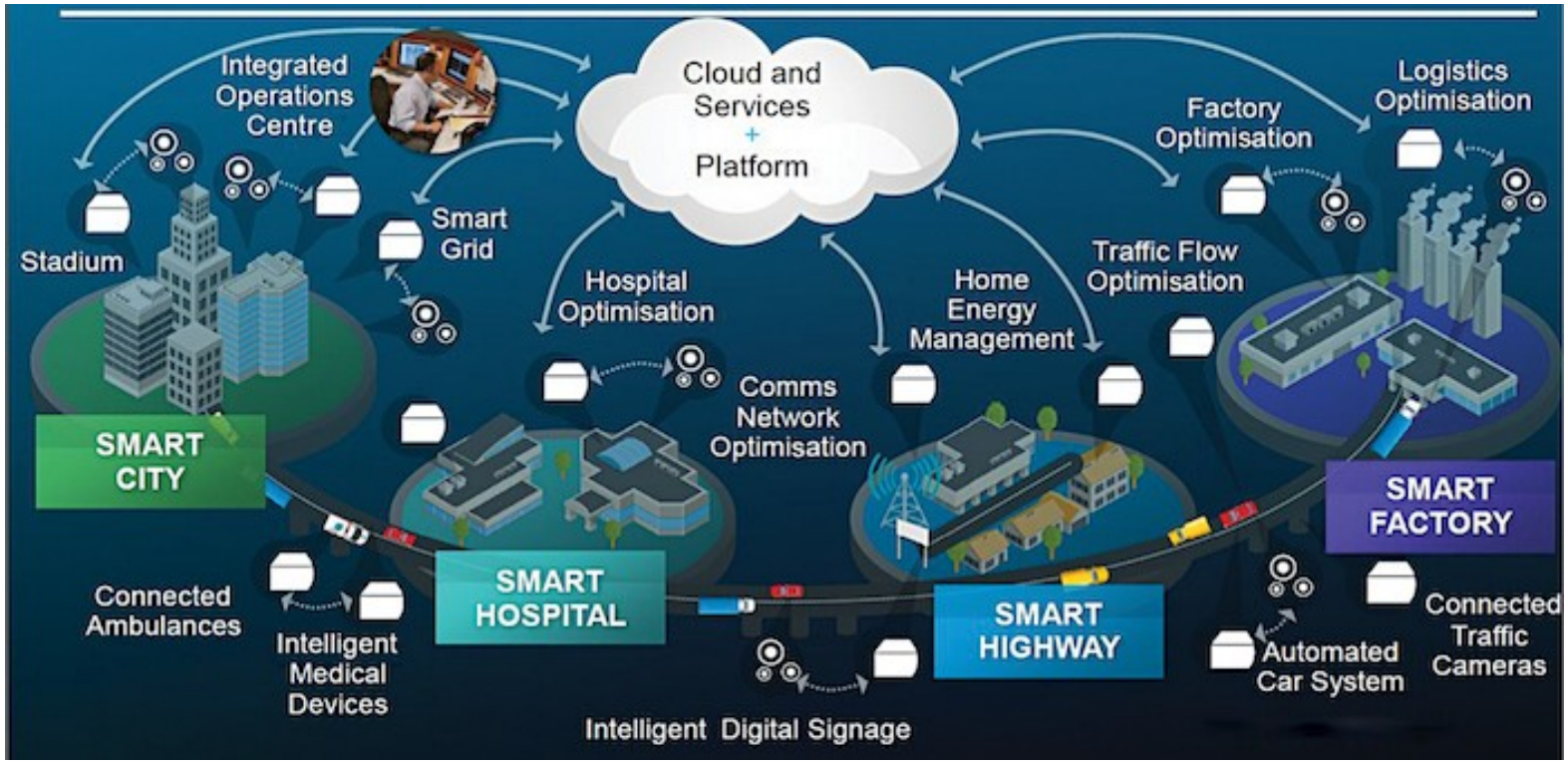
IZVORI PODATAKA

PAMETNA OKRUŽENJA

IoT značajno doprinosi Big Data izazovima

Do 2023 je 15 milijardi IoT uređaja instalirano širom sveta

Do 2030 procena je da će biti 29 milijardi IoT uređaja

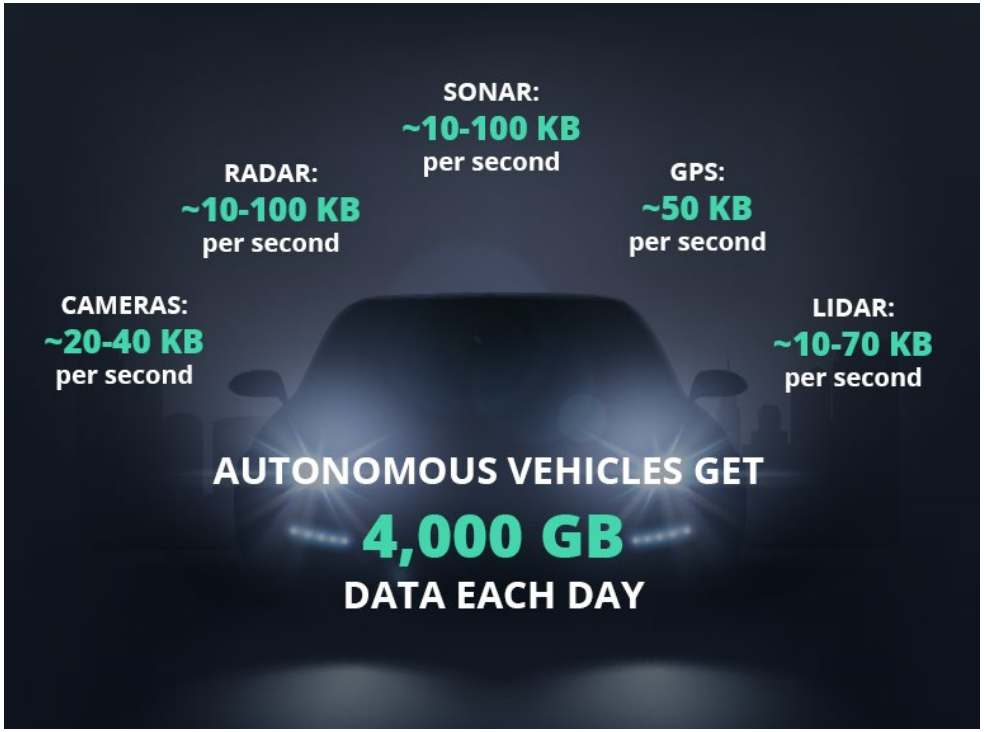


IIoT IZAZOV ZA BIG DATA

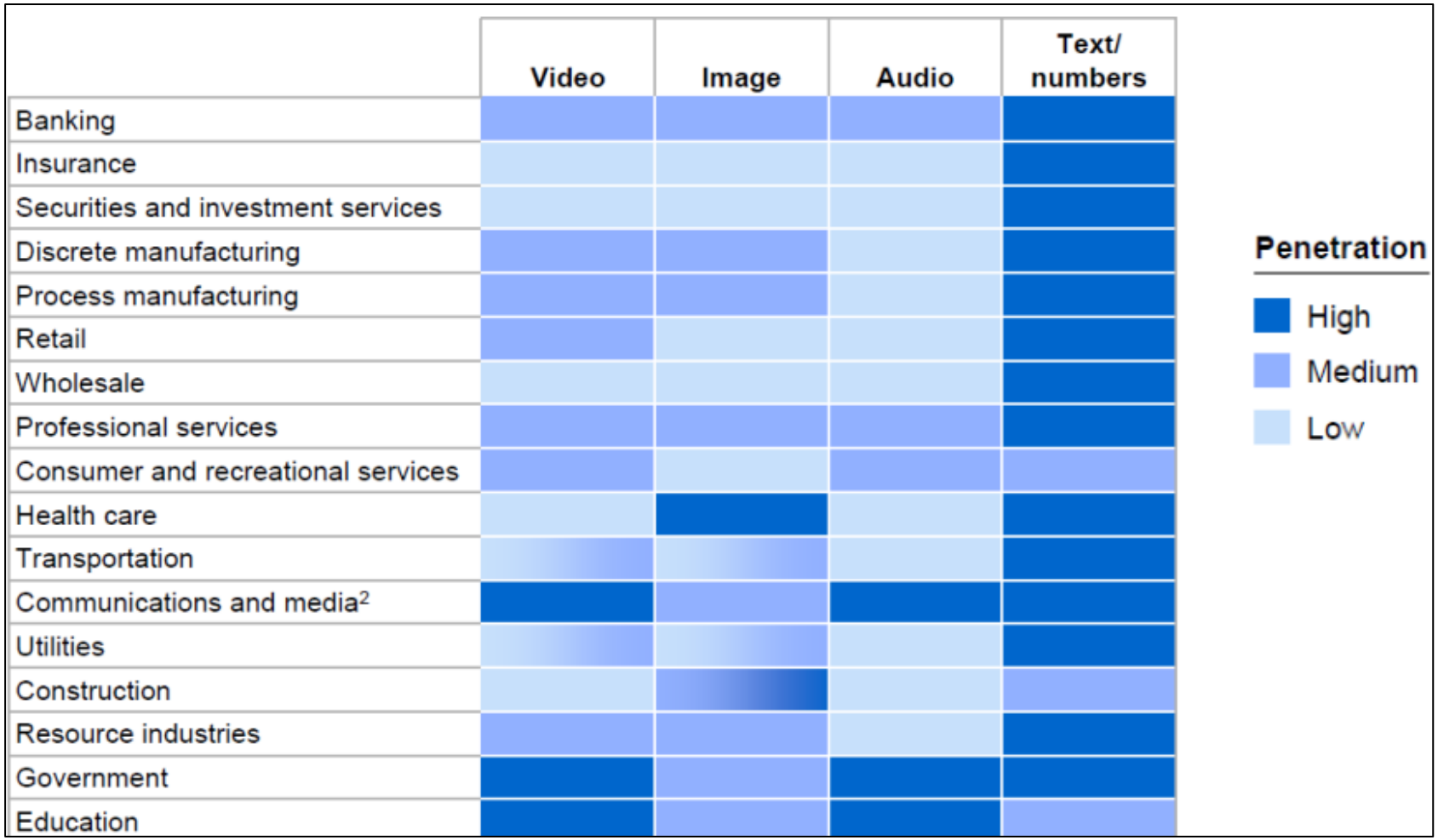
Industrial Internet of Things (IIoT)

Primer: Autonomna vozila

Samo jedno autonomno vozilo koristiće oko 4 TB podataka/dnevno

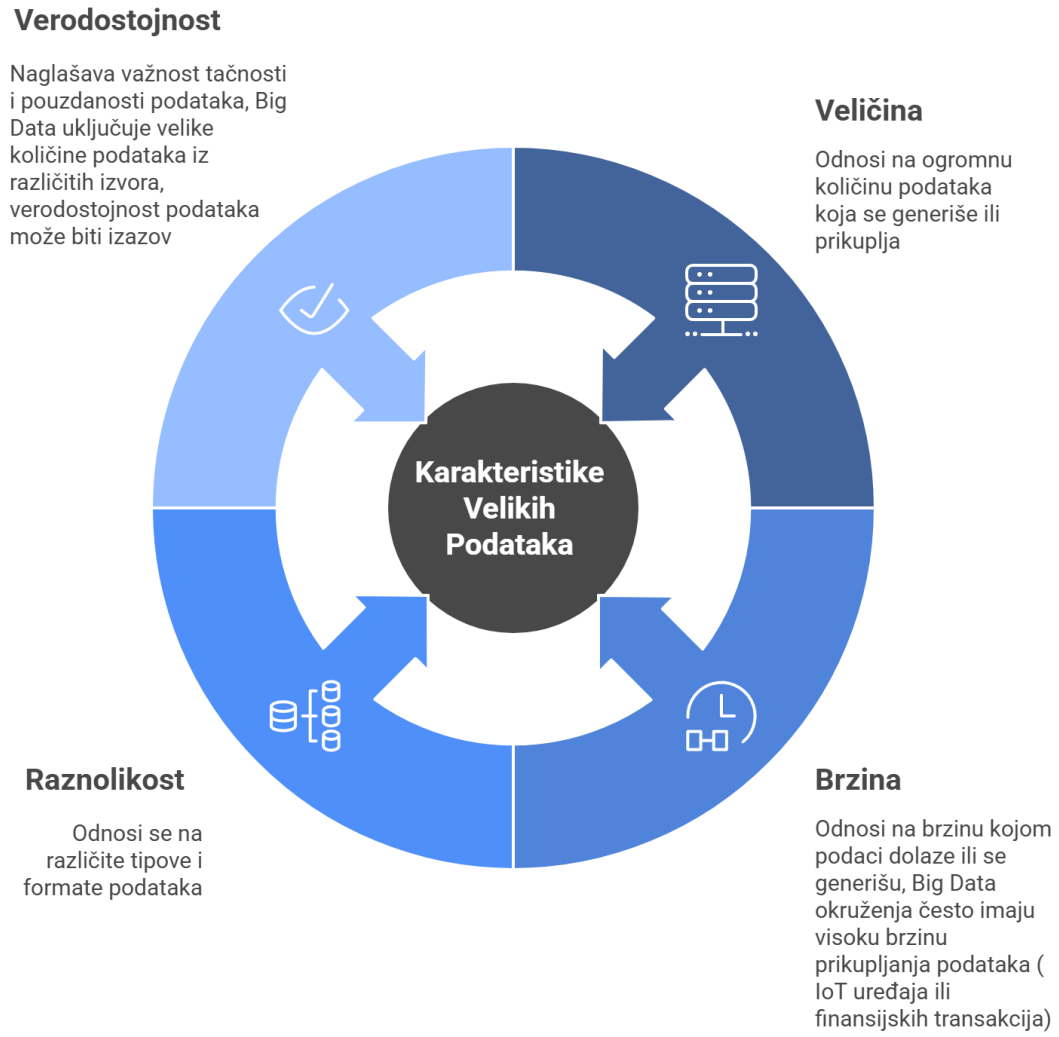


UČEŠĆE SEKTORA U BIG DATA

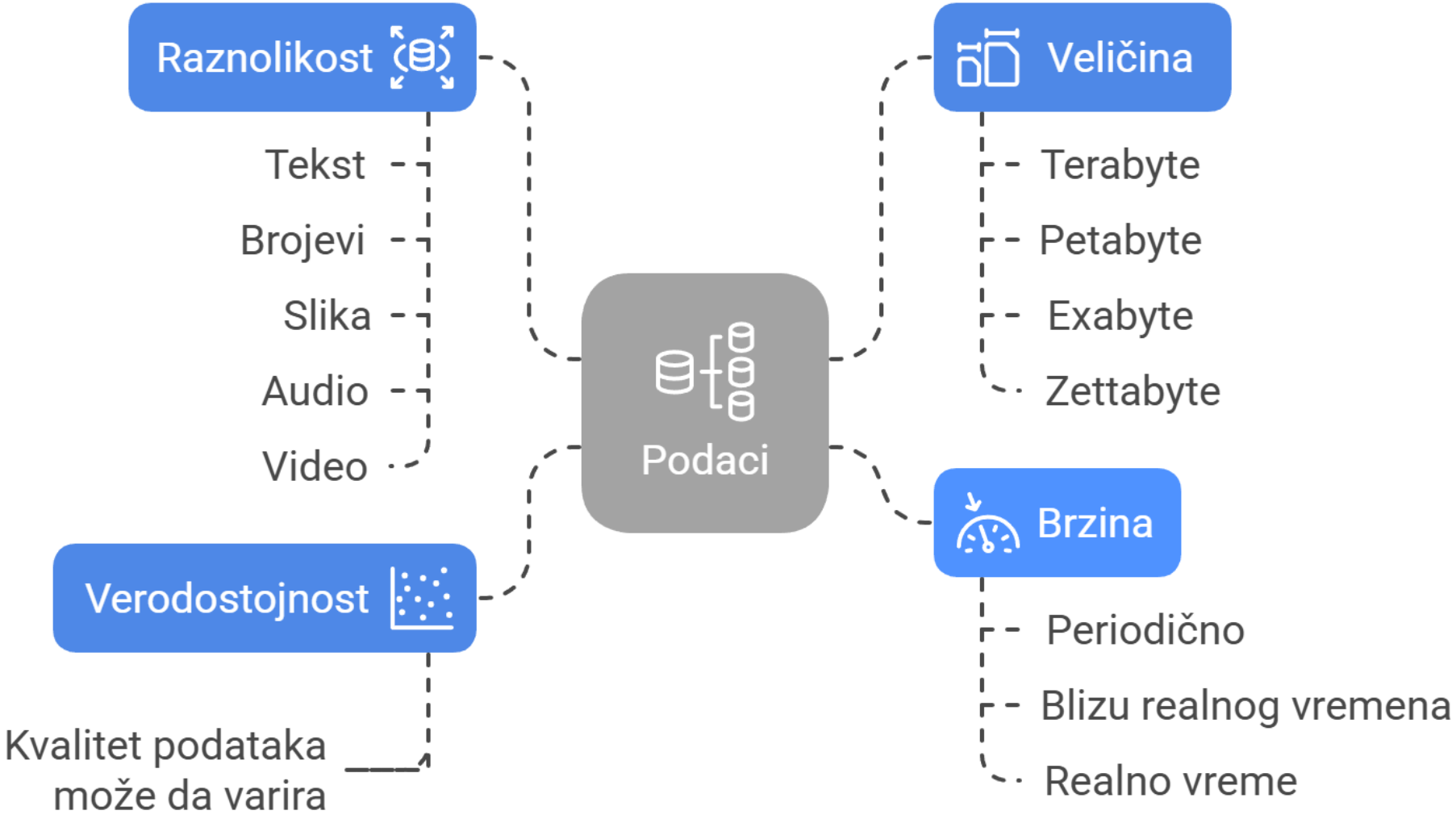


4V U BIG DATA

4V je termin koji se često koristi da bi se opisale ključne karakteristike podataka.



4V U BIG DATA


Raznolikost
Tekst
Brojevi
Slika
Audio
Video
Veličina
Terabyte
Petabyte
Exabyte
Zettabyte
Podaci
Verodostojnost
Kvalitet podataka može da varira
Brzina
Periodično
Blizu realnog vremena
Realno vreme

5V U BIG DATA

Ekstrakcija Vrednosti iz Velikih Skupova Podataka

ZAŠTO SADA

"Software is eating the world" , Marc Andreessen, The Wall Street Journal, 2011

"Data is the new oil", The Economist, 2017

40 godina baza podataka → volume

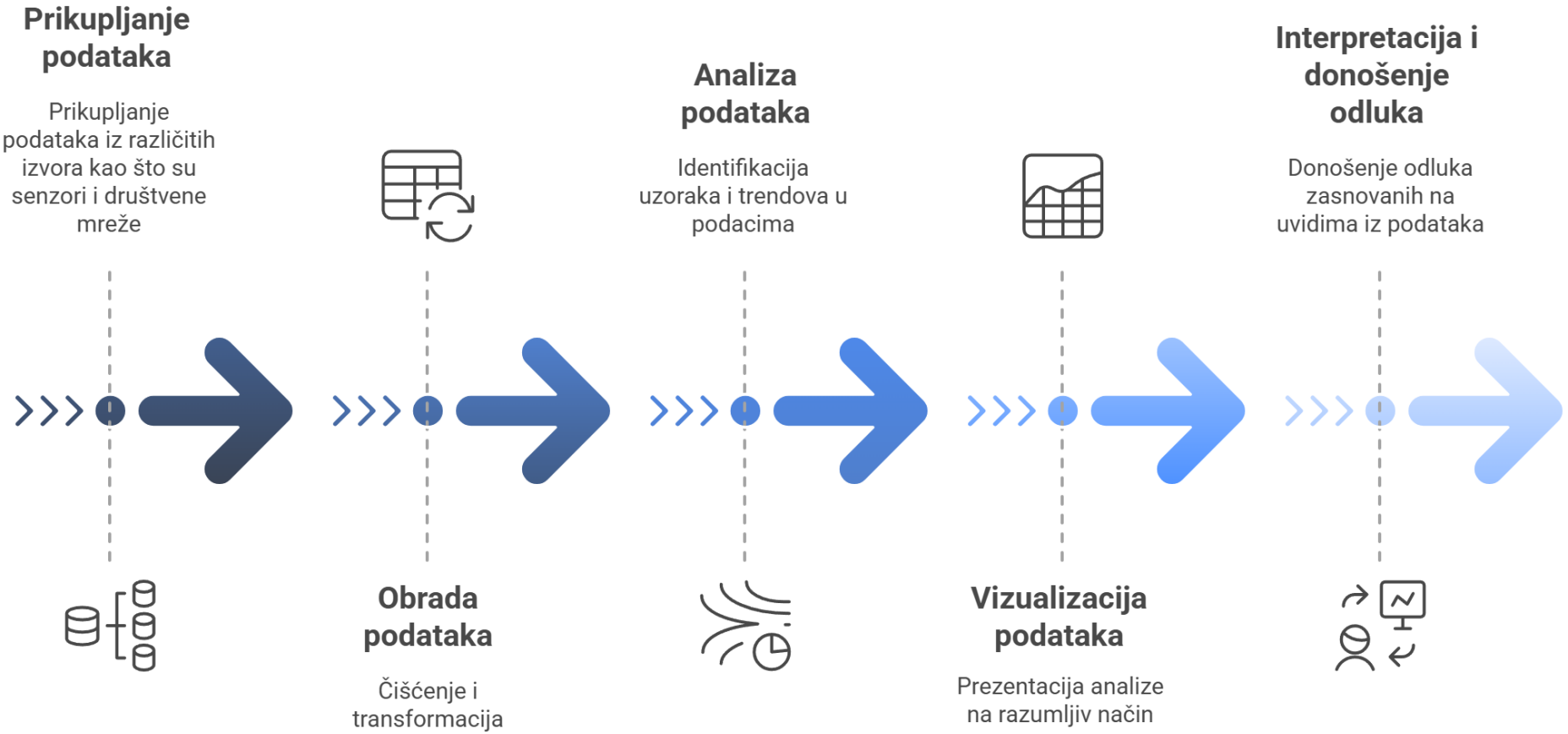
68% kompanija je investiralo u Big Data in 2018

57 milijardi $ je investirano u Big Data u 2018

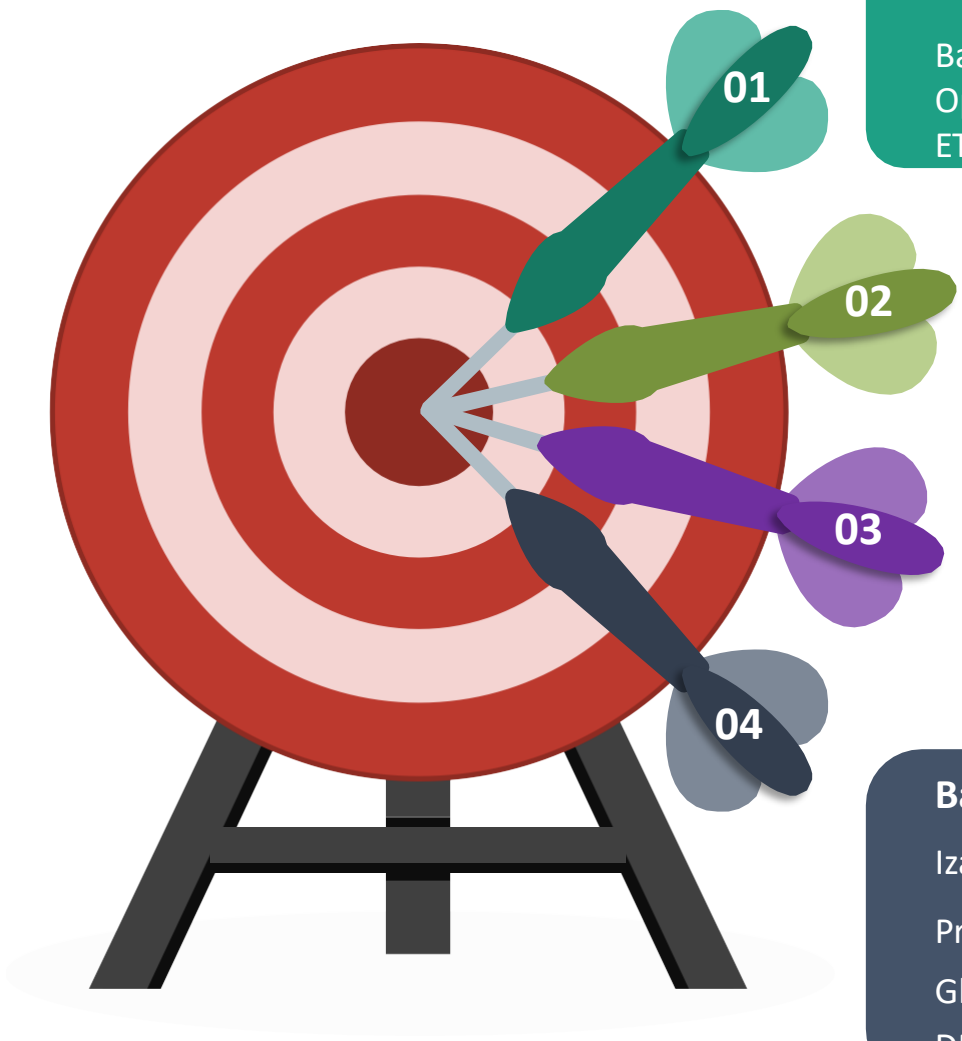
Large Hadron Collider (LHC) CERN je sakupljeno 200 Petabytes od 2012 a 1 Petabyte podataka se obrađuje svakog dana.

DNK (dezoksiribonukleinska kiselina) jednog pojedinca obično sadrži oko 3,2 milijarde baznih parova DNK.

PROCES UPRAVLJANJA VELIKIM PODACIMA



PODACI SU SVUDA OKO NAS



Terminologija

Baze podataka, DBMS, Šema,
Operational/Analytics Data Warehouse/Lake,
ETL/ELT, Skaliranje

Tehnologije

SQL DB, NoSQL DB, Distribuirane SQL DB
In-memory DB, Time-series DB
[Koncept + Prednosti + Primena]

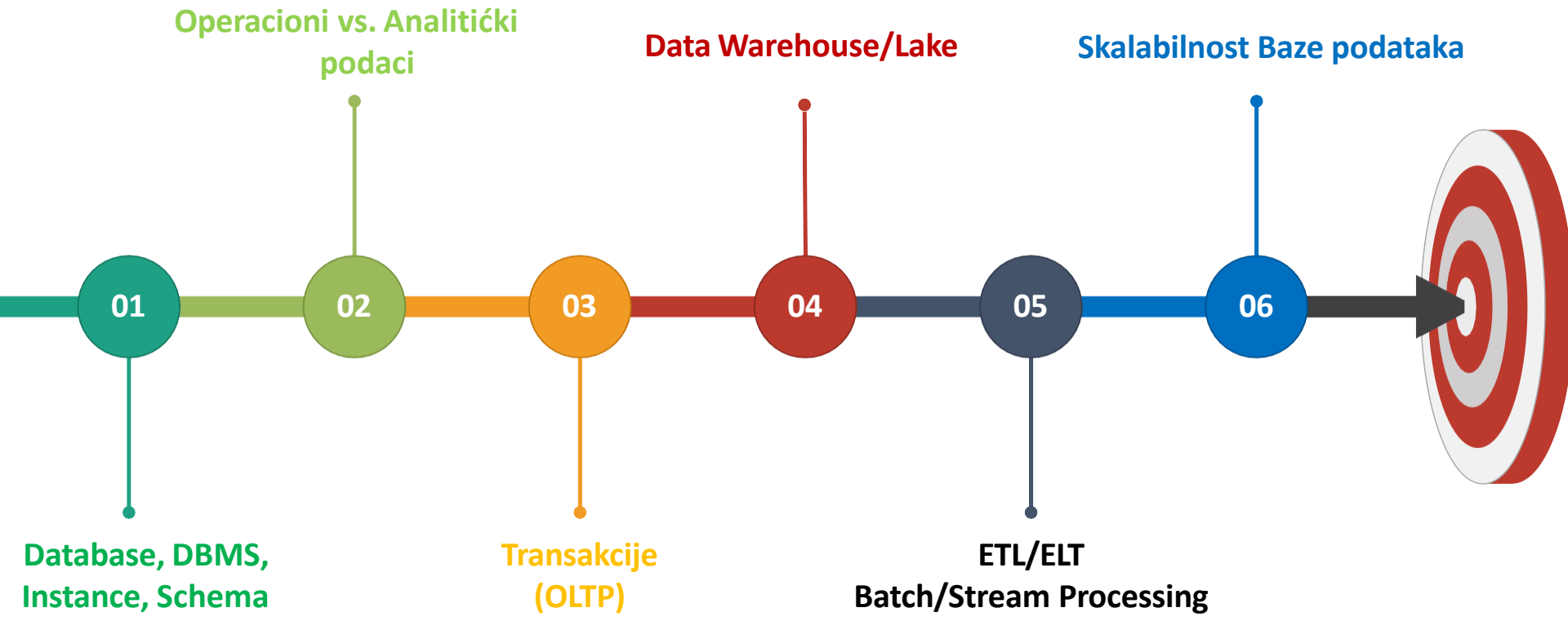
Vrste

#1 - Key-Value, #2 - Document
#3 - Wide Column, #4 - Graph

Baze podataka u oblaku (DBaaS)

Izazovi kod tradicionalnih baza podataka
Prelazak na DBaaS model
Glavne prednosti
DBaaS – Azure, AWS, GCP

Terminologija Baze podataka



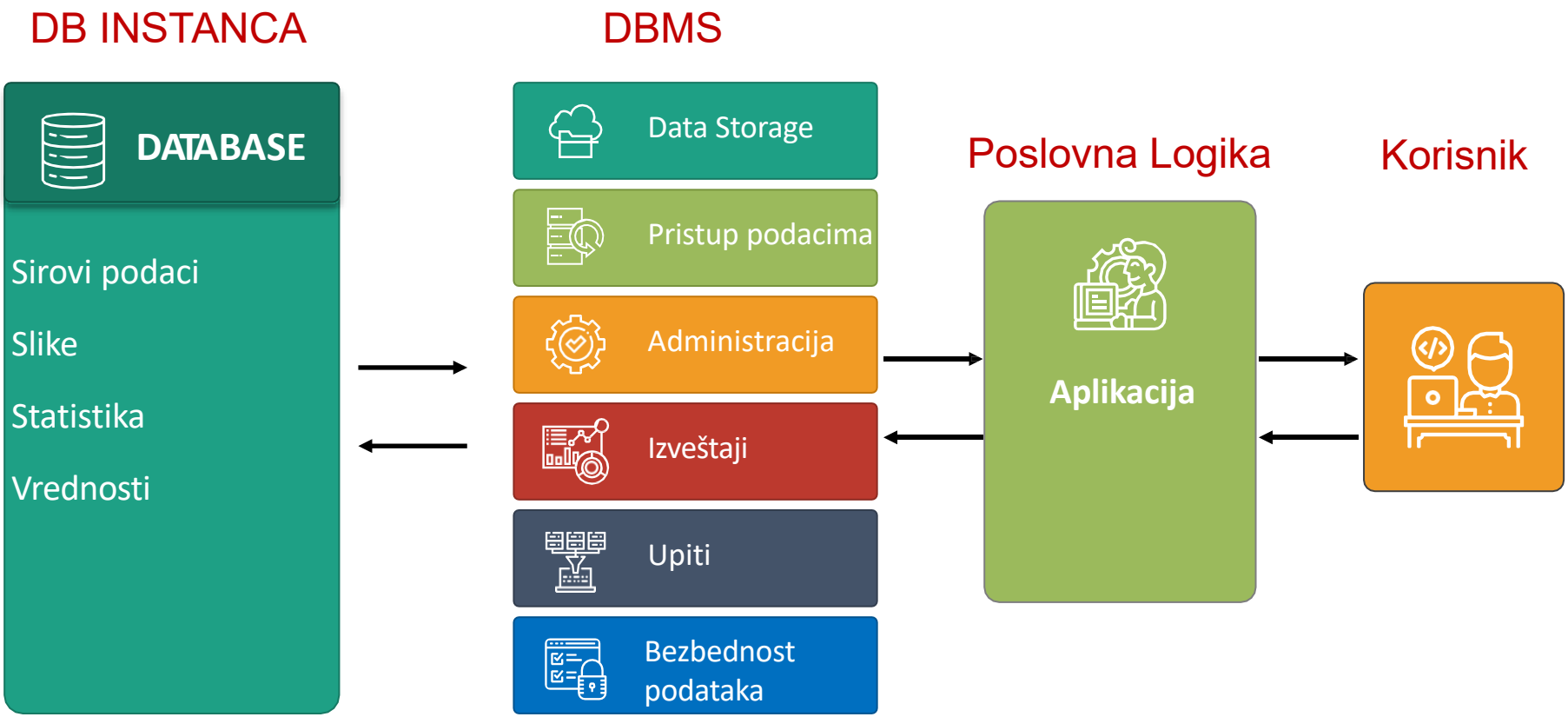
BAZA PODATAKA

Baza podataka je organizovana kolekcija podataka koja čuva podatke i kojoj se pristupa od strane aplikacije

Korisnički profili

Cene na tržištu (berzi)

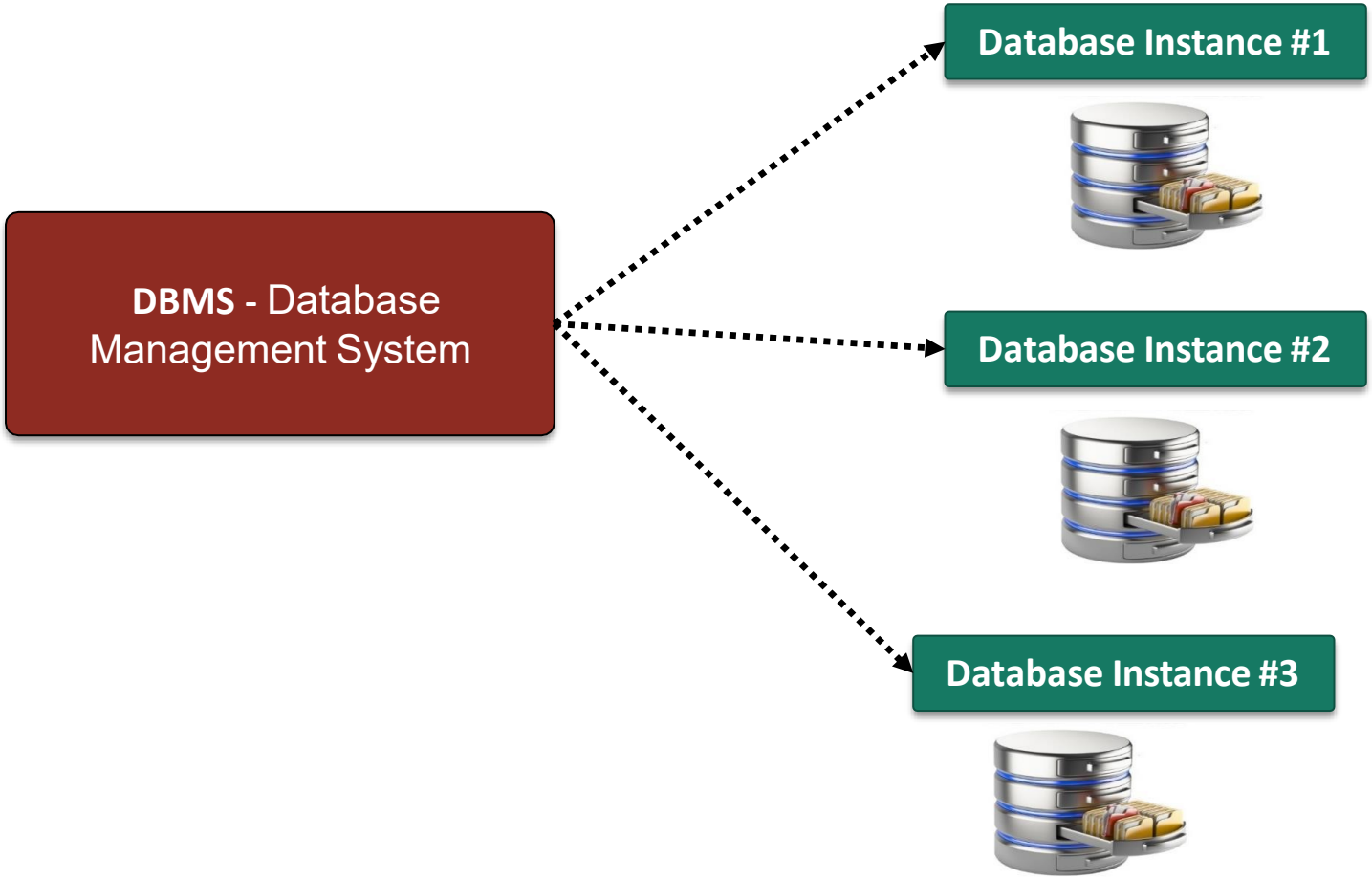
SISTEM ZA UPRAVLJANJE BAZOM PODATAKA



Instanca Baze podataka

Instanca baze podataka je logički entitet (kontejner) koji je kreiran od strane korisnika i spreman je da se popuni podacima

Kontejner sa podacima

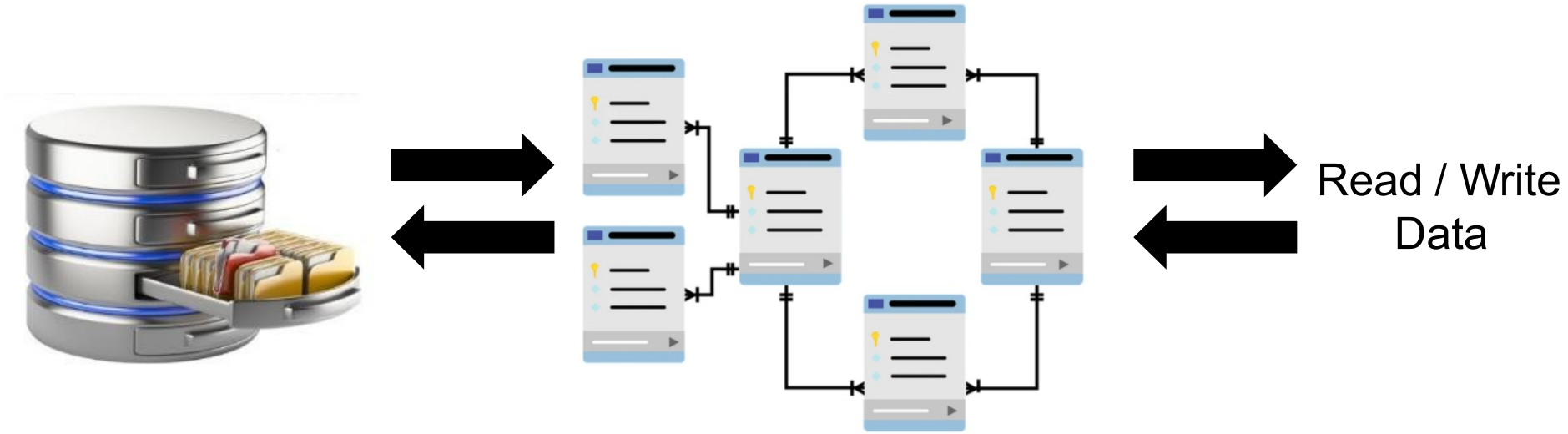


Šema Baze podataka

Šema baze podataka obezbeđuje da podaci budu organizovani u strukturu sa jasno definisanim setom pravila i ograničenjima koja su nametnuta od strane DBMS-a.

Šema baze podataka je šablon koji opisuje strukturu i koja se definiše nakon kreiranja instance baze podataka

Postoje DBMS koji se striktno drže definisane šeme ali postoje i baze koje su fleksibilnije



Tip Baze podataka

Tip baze podataka se bira na osnovu tipa podataka koji se čuva i na osnovu primene samih podataka (use case).

Dve osnovne vrste primene podataka u kompaniji su operacioni podaci i podaci za analizu.

Operacioni podaci su svakodnevni podaci koji se kreiraju na dnevnom nivou u realnom vremenu primenom transakcija

Analitički podaci i operacioni podaci se drugačije skladište i obrađuju u bazama podataka

Operacioni Podaci

- Porudžbine
- Proizvodi
- Klijenti/Snabdevači
- Servisi
- Tiketi
- Prodaja

Analitički Podaci

OLTP (Online Transactional Processing)

Transakcije

Transakcija u bazama podataka je jedinica posla koja se izvrši unutar baze

Primer je plaćanje robe gde je transakcijom obuhvaćeno skidanje novca sa računa upatioca, prebacivanje novca na račun primaoca i ažuriranje brojčanog stanja kupljenog artikla.

Podaci se čuvaju u dvodimenzionalnim strukturama

Transakcije obrađuju sistemi poznati kao OLTP (ERP, CRM, Plaćanja, ...) aplikacije

OLTP aplikacije koriste OLTP bazu podataka tj bazu podataka koja podržava transakcije

ACID Model (Atomic, Consistent, Isolation, and Durable)

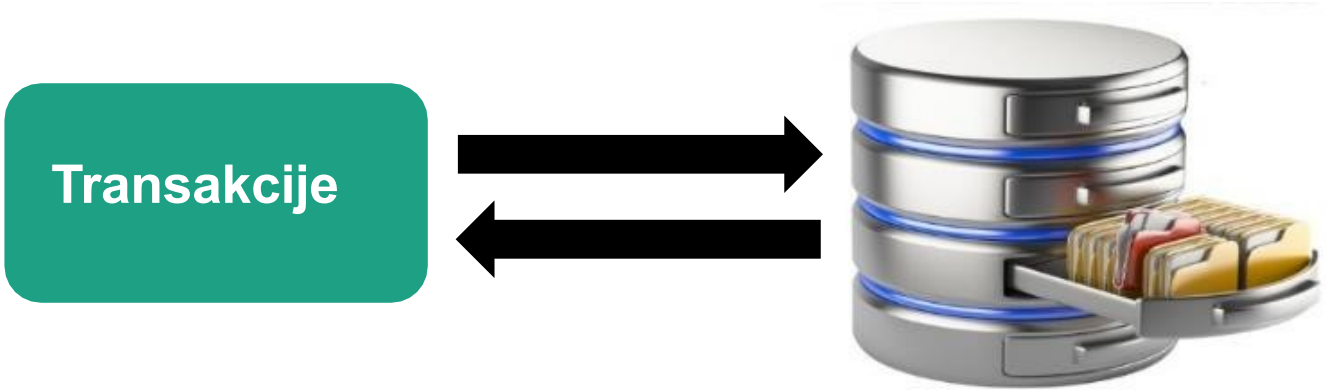
Uzimanje novčanica sa računa

Transakcija
Confirmed/Not Confirmed

OLTP (Online Transactional Processing)

Transakcije u OLTP sistemima

- Sve aktivnosti se čuvaju u bazu podataka preko transakcija
- Obrada velikog broja transakcija u realnom vremenu **(High-volume)**
- Brz pristup podacima u bazi **(Low-latency)**
- Česte promene podataka (updated transakcije)



OLAP (Online Analitical Processing)

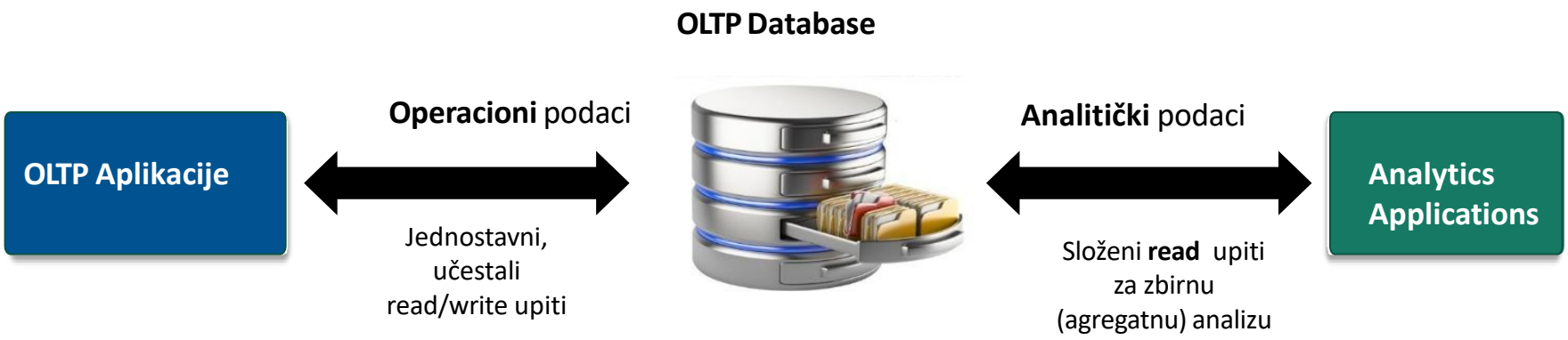
- **Analitički podaci**
  - Poslovne odluke na osnovu analize istorijskih podataka
  - Ulaz za *Business Intelligence* (BI) sisteme
  - Analitički podaci su operacioni podaci ali u mnogo većem obimu
  - Kompleksna analiza
    - Utiče na bolje poslovanje
    - Fokus na novim prilikama
    - Prepozna trend na tržištu
    - Prepoznaju anomalije

OLAP (Online Analitical Processing)

Opcija #1 – Ista OLTP Baza podataka za transakcije i analizu podataka

OLTP baza podataka nije dizajnirana da obrađuje veliku količinu agregiranih podataka kroz kompleksne upite koje bi uticale na performanse sistema

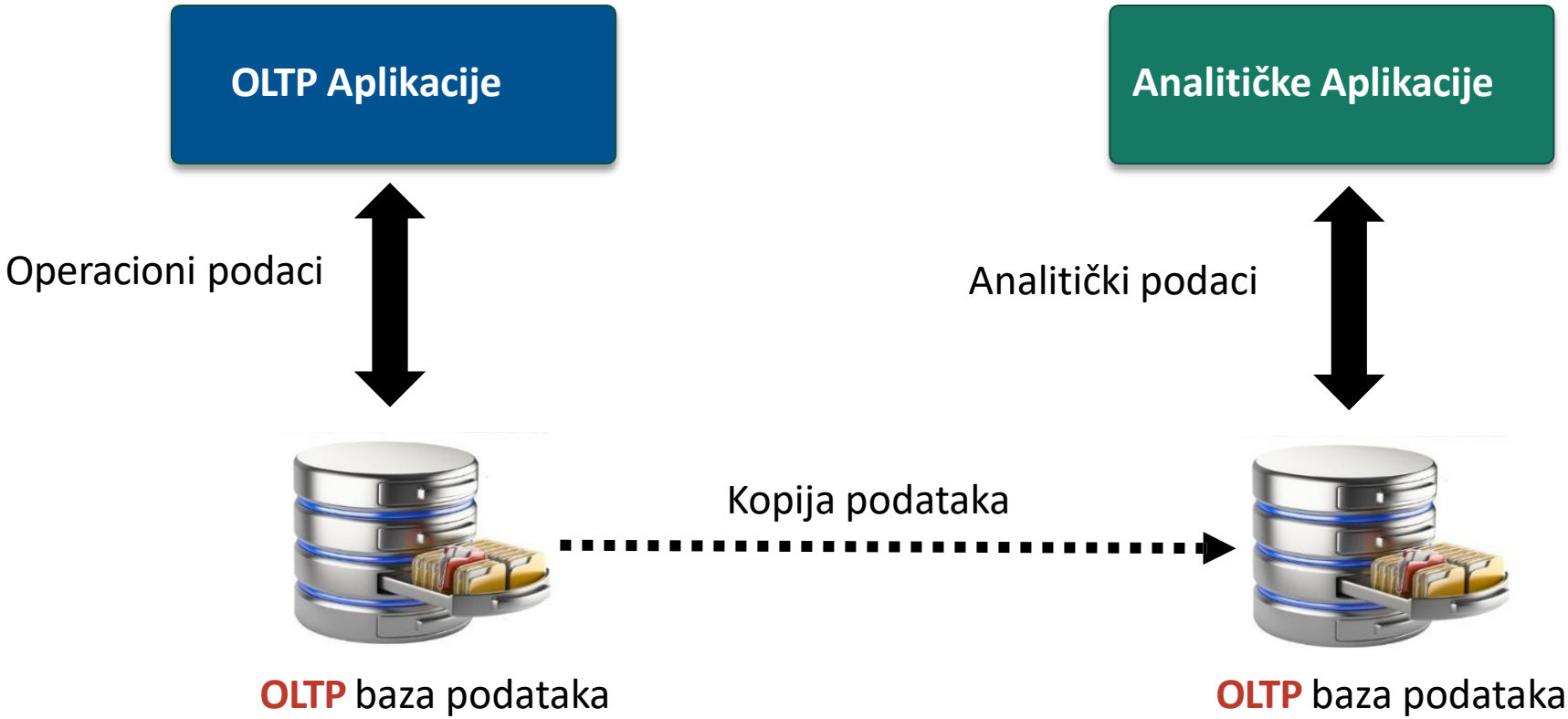
OLAP je tehnologija baze podataka je optimizovana za upite i izveštavanje, umesto obrade transakcija



OLAP (Online Analitical Processing)

Opcija #2 – Dedicated OLTP Baza podataka samo za analizu podataka

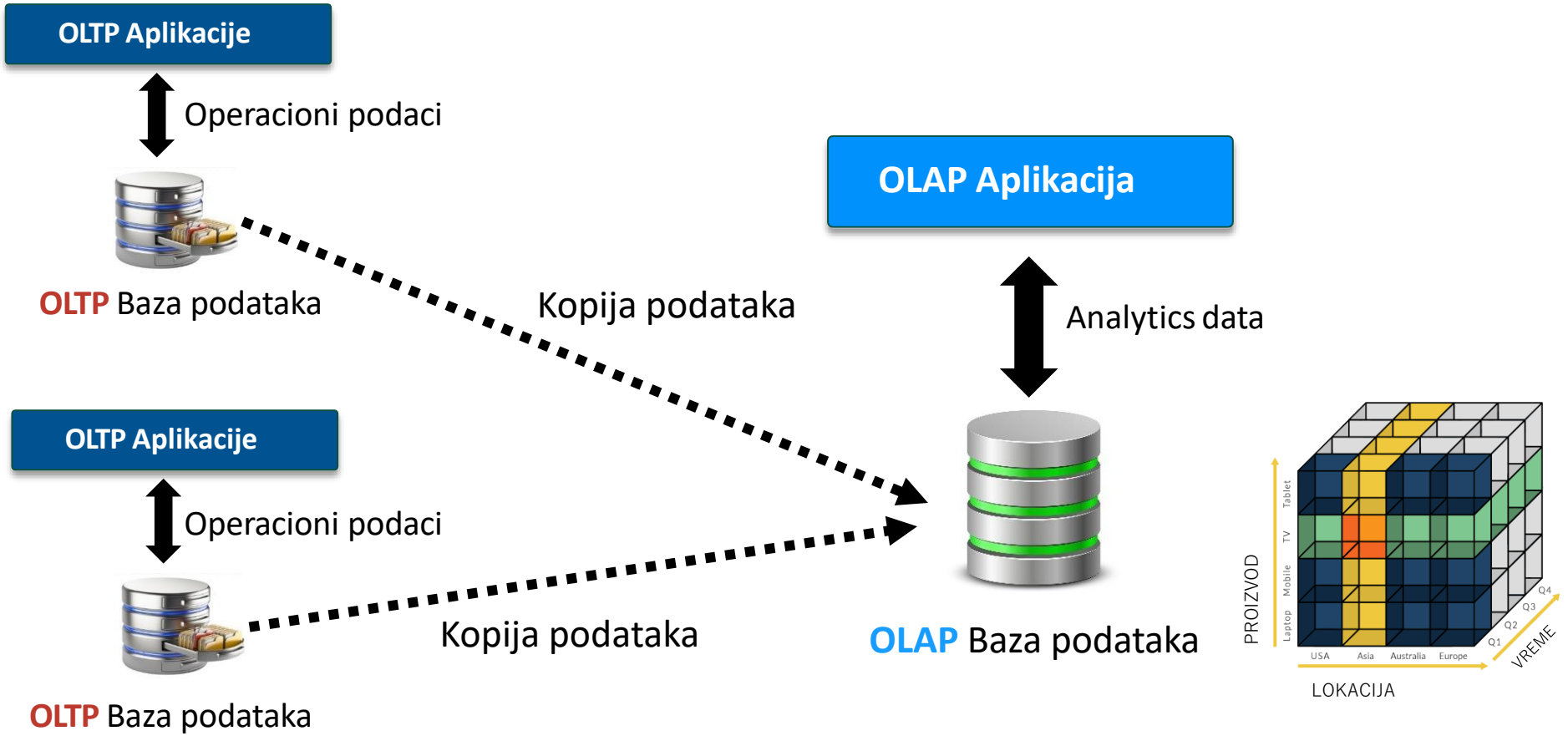
Ovakva arhitektura za analizu podataka radi racionalno samo za jednostavne analitičke zahteve



OLAP (Online Analitical Processing)

Opcija #3 – OLAP

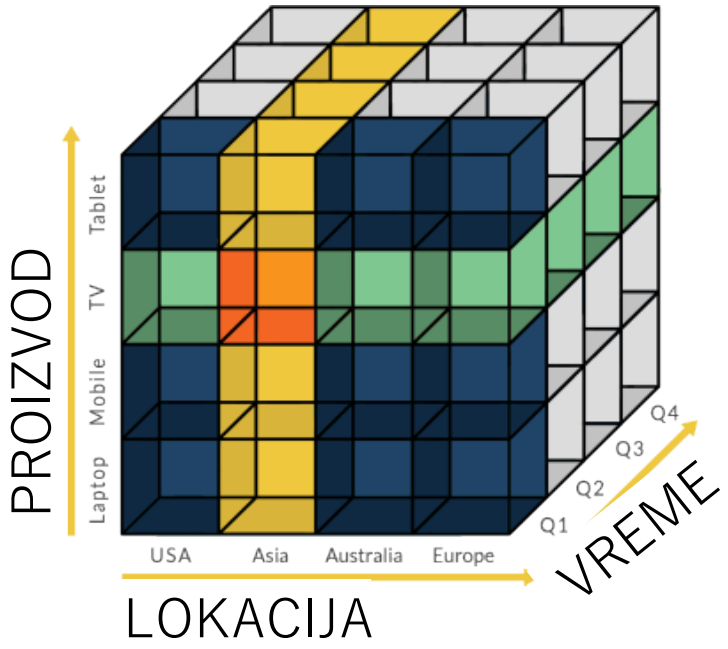
Podaci se čuvaju u drugačijoj strukturi u odnosu na OLTP bazu podataka (dvodimenzionalna) koja je optimizovana za analitičke upite – višedimenzionalna struktura podataka Kocka



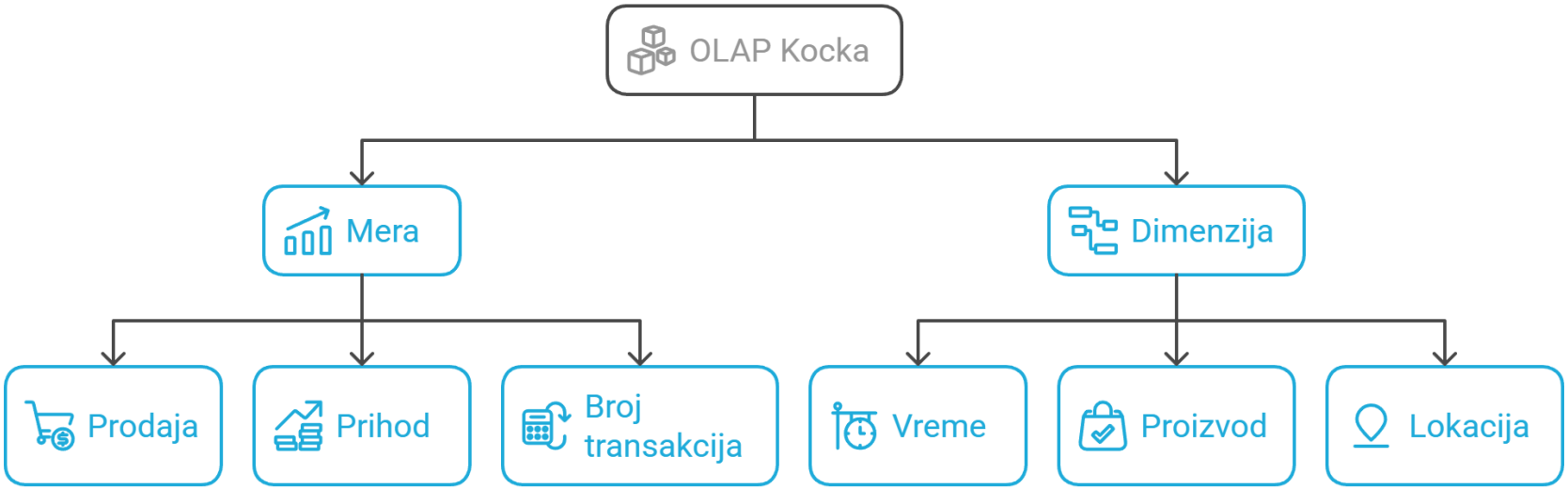
VIŠEDIMENZIONALNOST

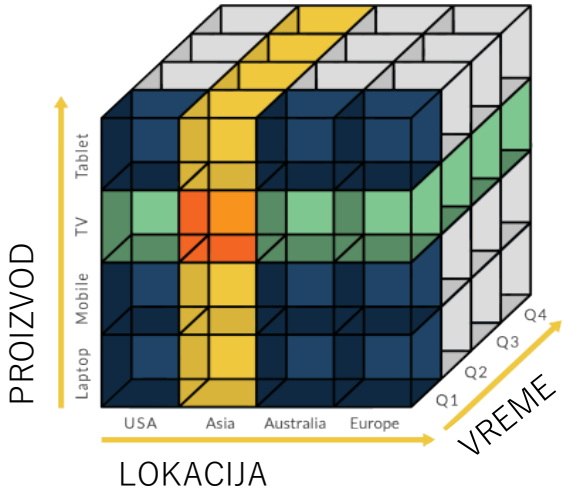
Opcija #3 – OLAP

- Višedimenzionalnost je osnovna karakteristika OLAP baze podataka
- Korisnik ima mogućnost da parametre poslovanja vidi u preseku dimenzija koje opisuju te parametre
- Ljudima je prirodno da podatke analiziraju kroz dimenzije
- Ako se pojava prati u tri dimenzije, reč je o kocki, a u više dimenzija o hiperkocki.

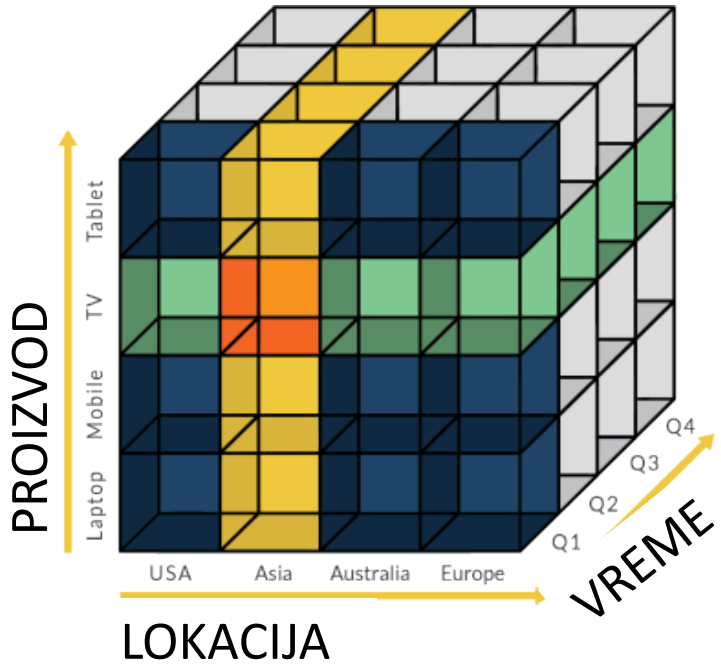


OLAP KOCKA – KLJUČNI POJMOVI





OPERACIJE NAD OLAP KOCKOM



Slice — Izbor jedne vrednosti u dimenziji | Pogledaj podatke za januar

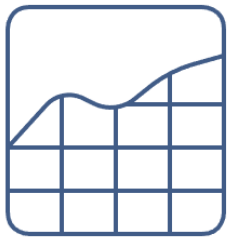
Dice — Izbor podskupa u više dimenzija | Pogledaj prodaju za 3 proizvoda u 2 regiona

Drill-down — Ideš u detalje (veća rezolucija) | Sa godine na mesece

Roll-up — Grupisanje u veće agregate | Sa meseci na kvartale

Pivot — Promena prikaza dimenzija | Zameni redove i kolone u izveštaju

PRIMENA OLAP KOCKE



BI alati

Alati koji se koriste za analizu poslovne inteligencije.

Rešenja za Data Warehouse

Rešenja za upravljanje i analizu podataka.

Finansijska i prodajna analitika

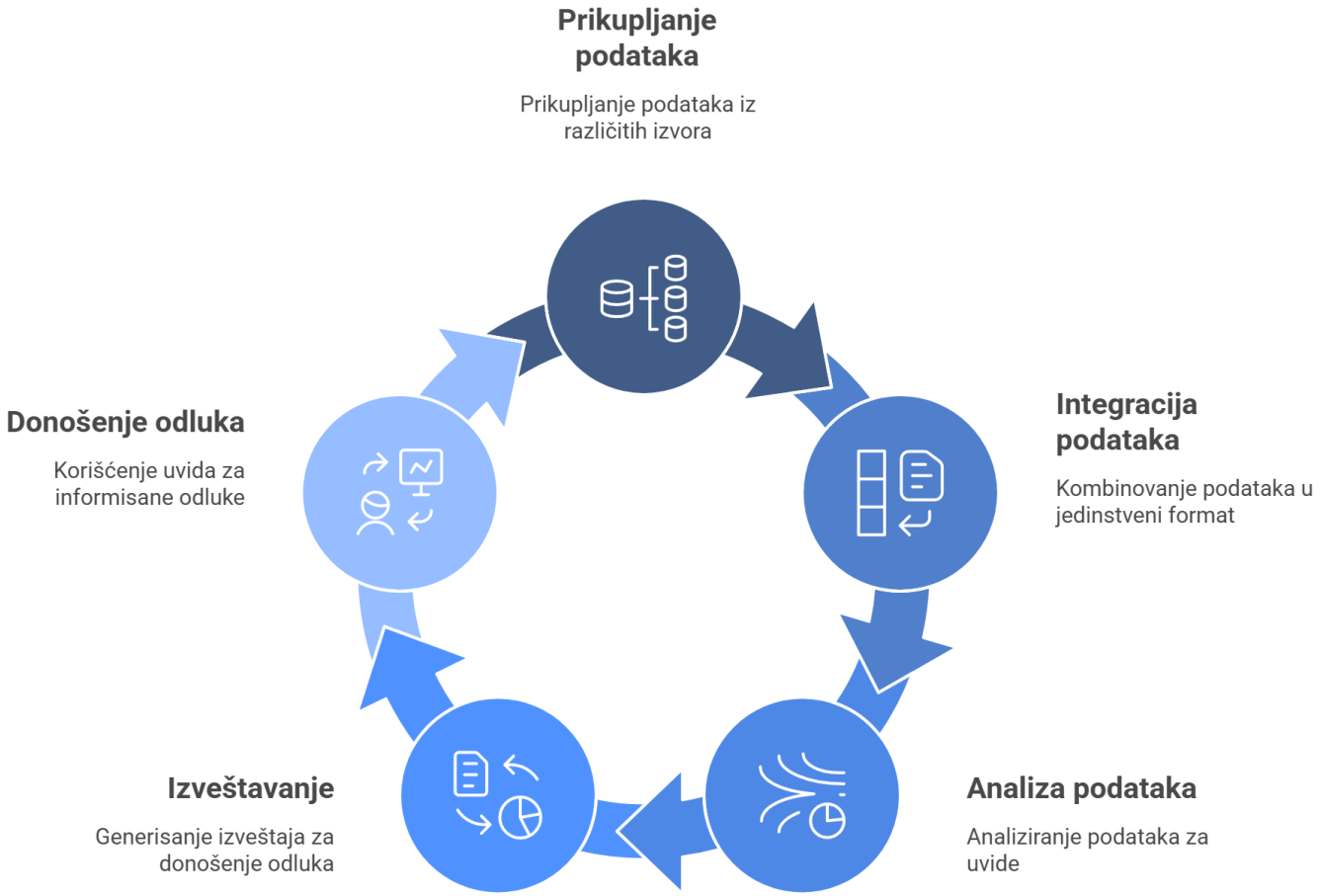
Analiza fokusirana na finansijske i prodajne podatke.

Planiranje i budžetiranje

Procesi za finansijsko planiranje i upravljanje budžetom.

Data Warehouse

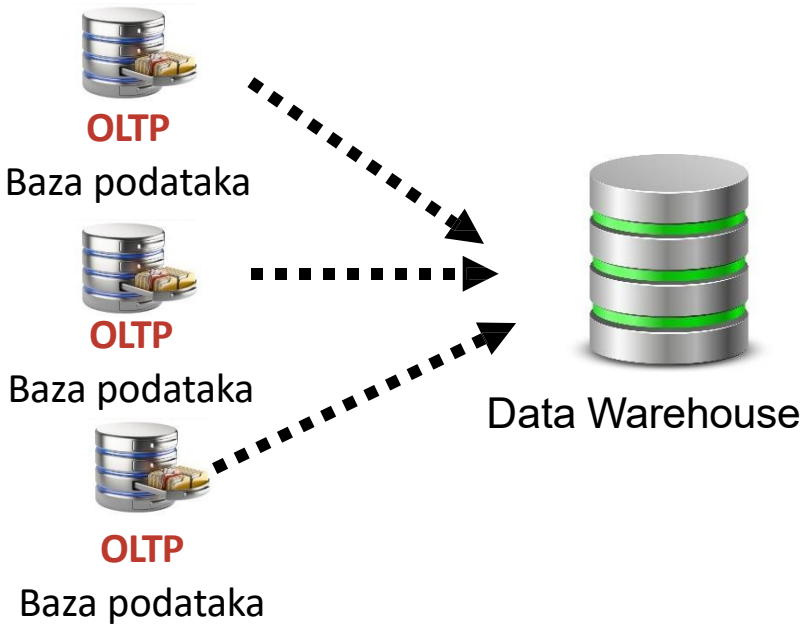
Data warehouse (skladište podataka) je centralizovani sistem za skladištenje, integraciju i analizu velike količine podataka iz različitih izvora, koji se koristi za izveštavanje, analitiku i donošenje poslovnih odluka.

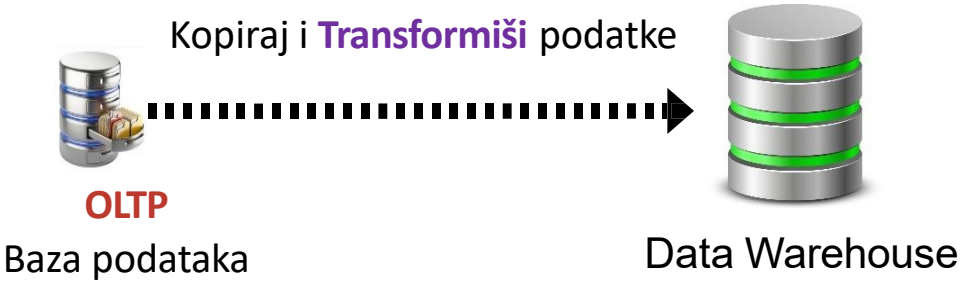


Ključne karakteristike Data Warehouse sistema

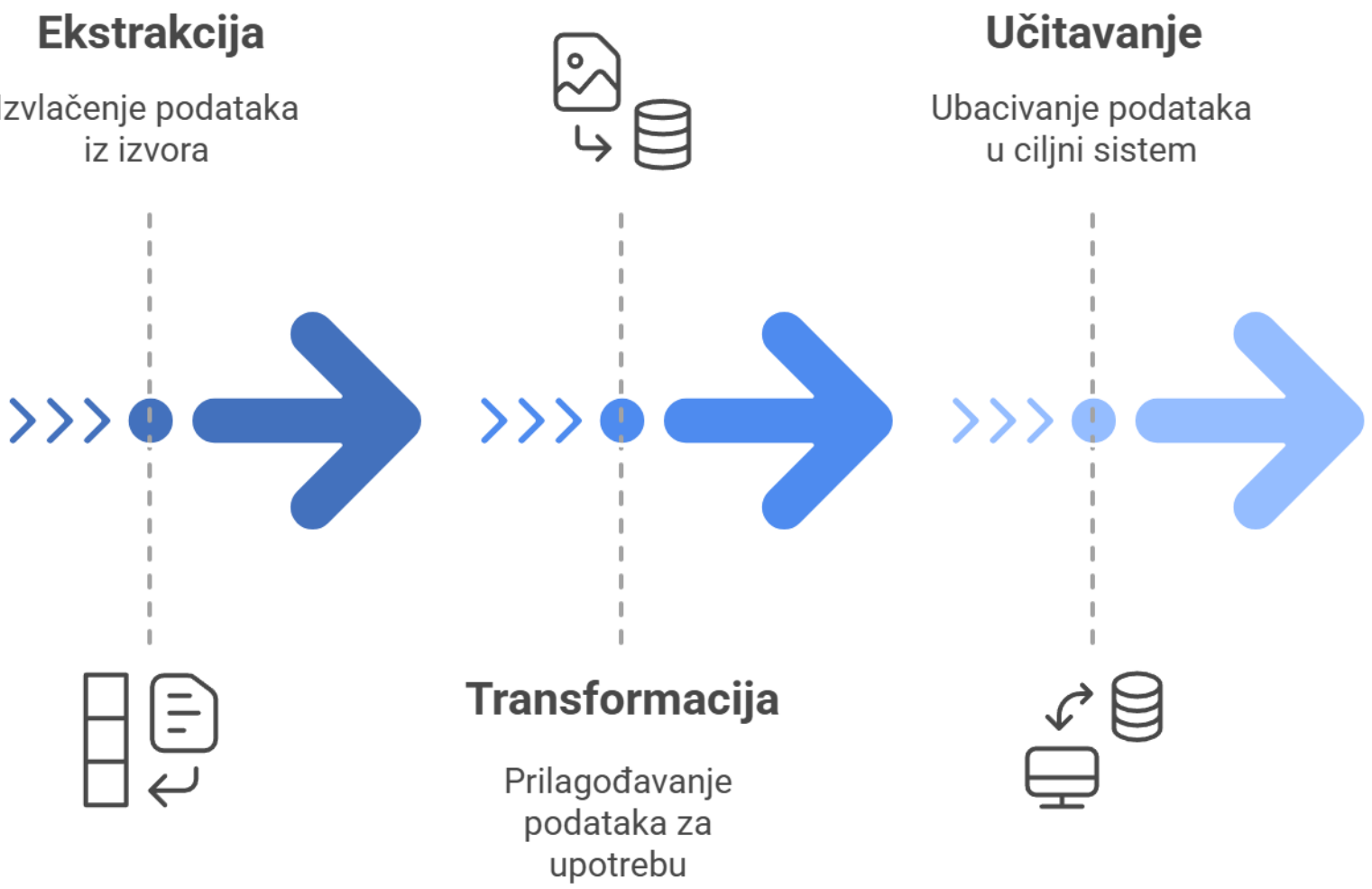
Data Warehouse

- Prikuplja podatke na jednom mestu sa različitih izvora podataka
- Radi sa velikom količinom podataka
- Optimizovan za OLAP aplikacije (kompleksne read upite)
- Čuva istorijske podatke na duže vreme (količina je u terabajtima i petabajtima)
- Sirovi podaci se prebacuju (transformišu) u novu strukturu optimizovanu za analizu podataka



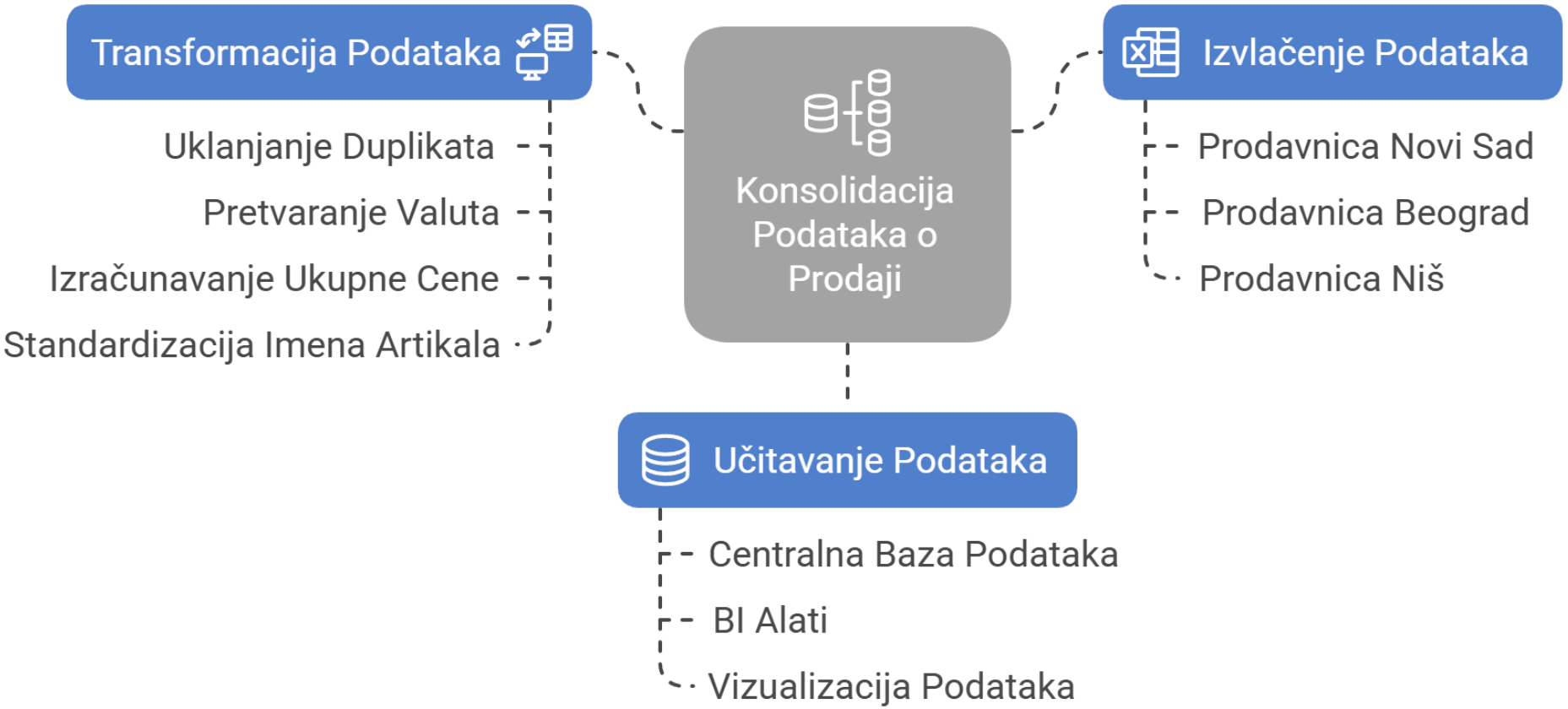


Premeštanje podataka između sistema OLTP - OLAP



Premeštanje podataka između sistema OLTP – OLAP

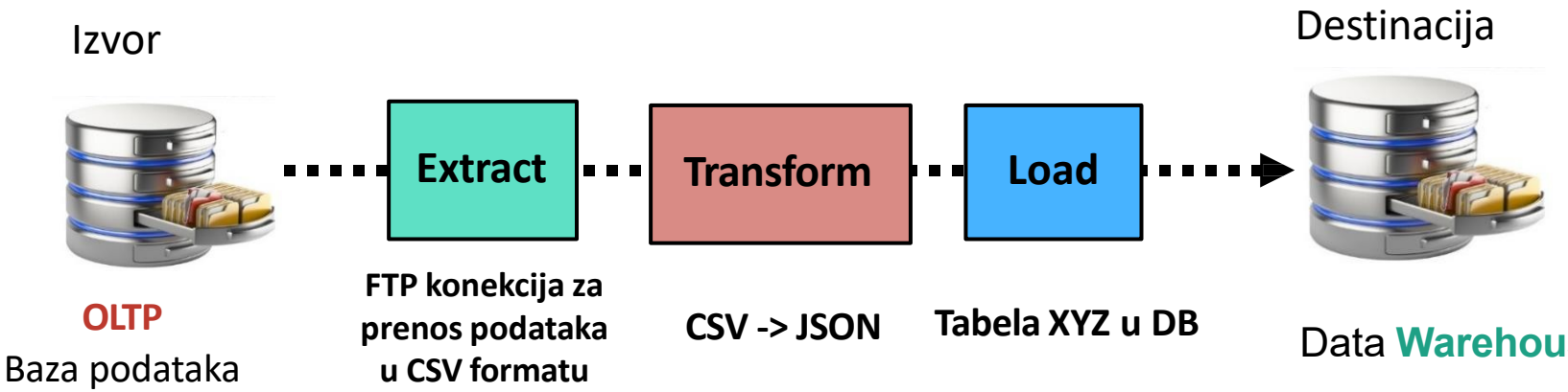
Primer 1



Premeštanje podataka između sistema OLTP – OLAP

Primer 2

- **ETL - Extract, Transform, Load**
  - Proces uzimanja (**extracting**) podataka iz jednog sistema (izvor podataka), **transformacija** podataka u novu strukturu, i unos (**load**) podataka u odredišni sistem
  - Podaci se transformišu pre nego da se unesu u odredišni sistem
  - Tipičan primer je Data Warehouse
  - **Izvorišna aplikacija** radi transformaciju podataka

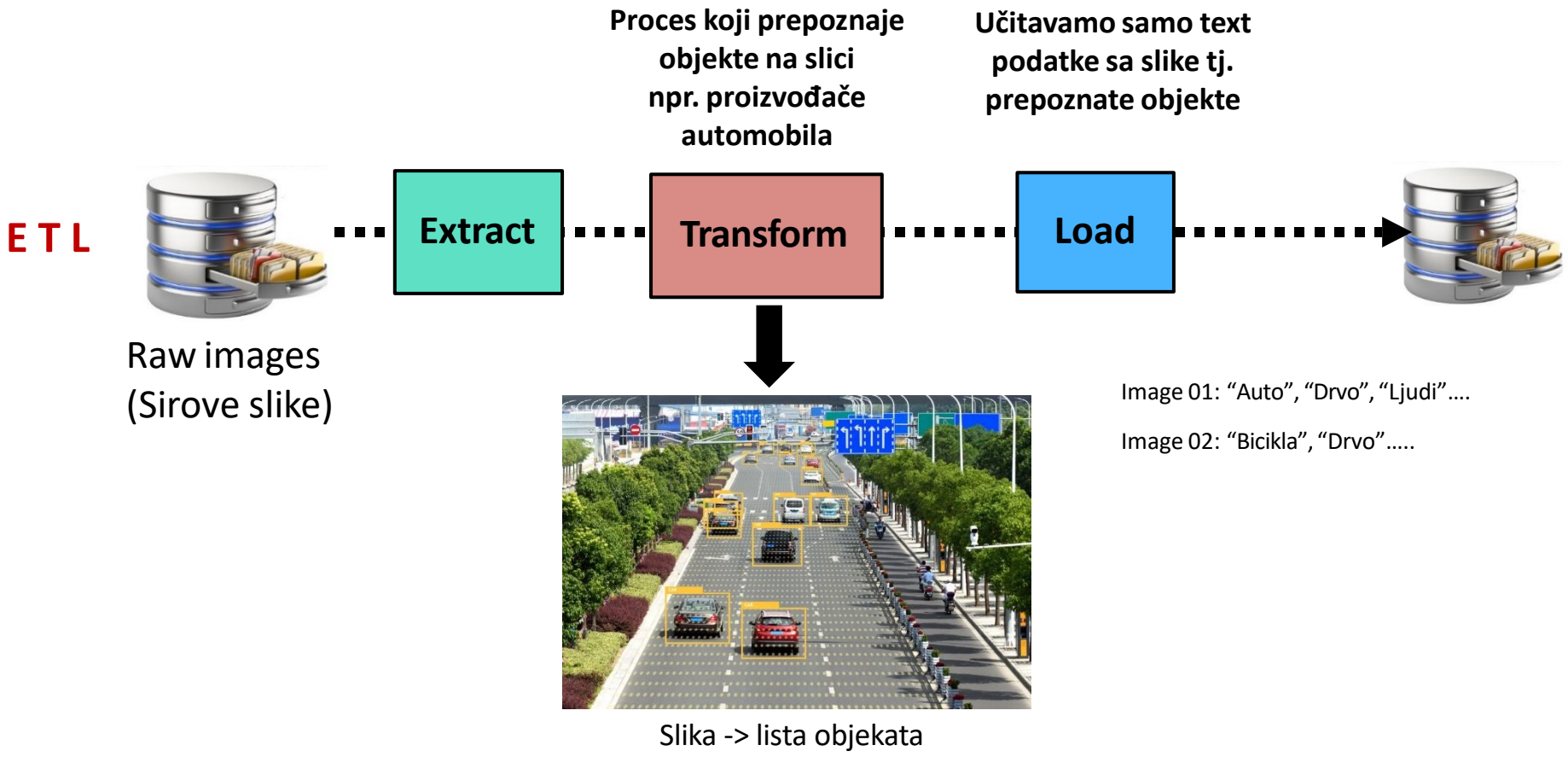


Premeštanje podataka između sistema OLTP – OLAP

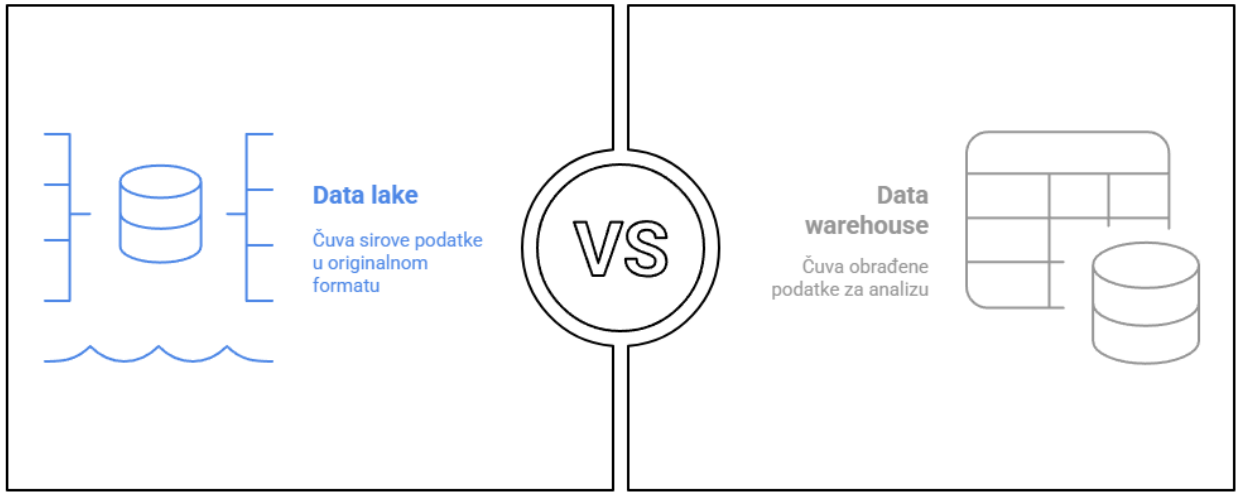
Primer 3

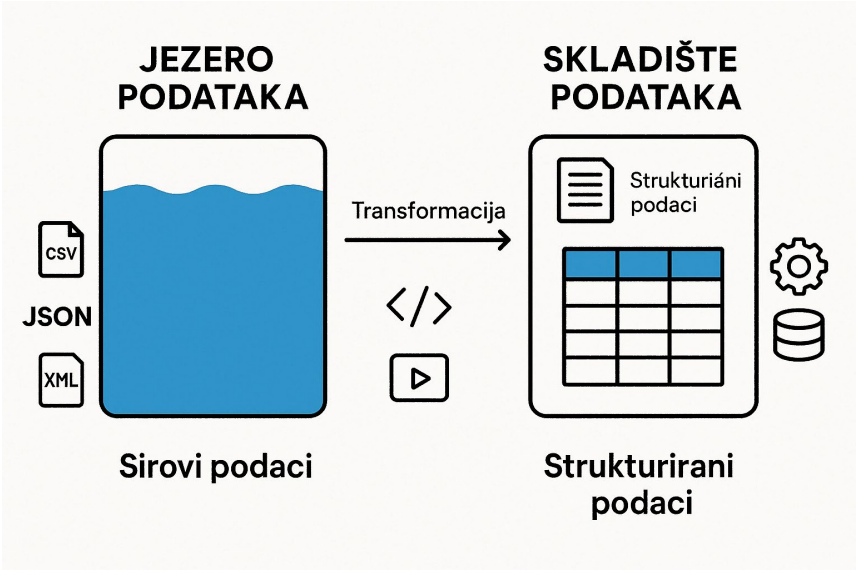
- App X – Prepoznaje listu objekata na slici

Efikasan metod za **umanjenje količine podataka** koji se čuva



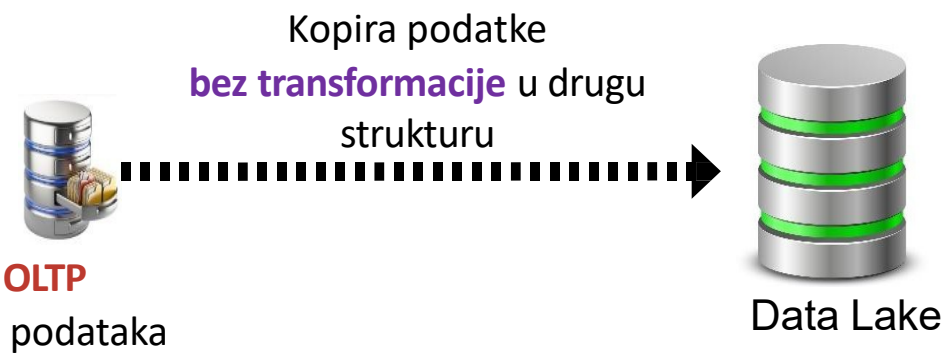
Data Lake vs Data Warehouse





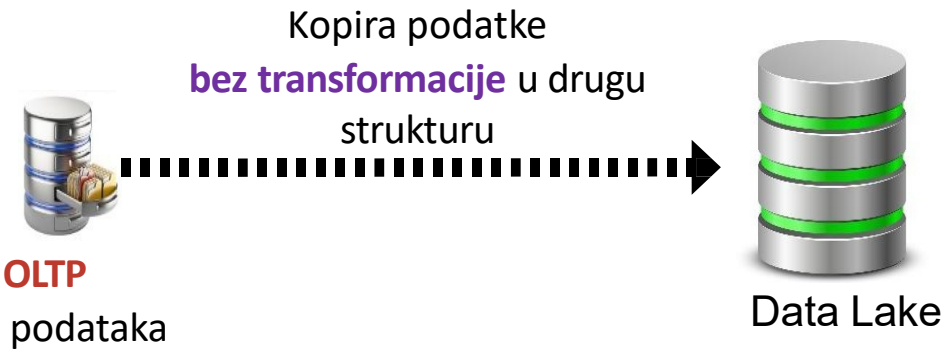
Osobine Data Lake Sistema

Sirovi podaci	Različiti formati	Skalabilnost	Kasnija obrada	Primeri platformi
Podaci se čuvaju bez prethodne obrade	Može sadržati različite tipove fajlova	Lako se proširuje	Analiza se radi po potrebi	AWS S3, Azure, Hadoop, GCP
[Skladištenje neobrađenih podataka]	[Podržava CSV, JSON, itd.]	[Proširuje se na petabajte]	[Šema se primenjuje prilikom čitanja]	[Popularne cloud platforme]

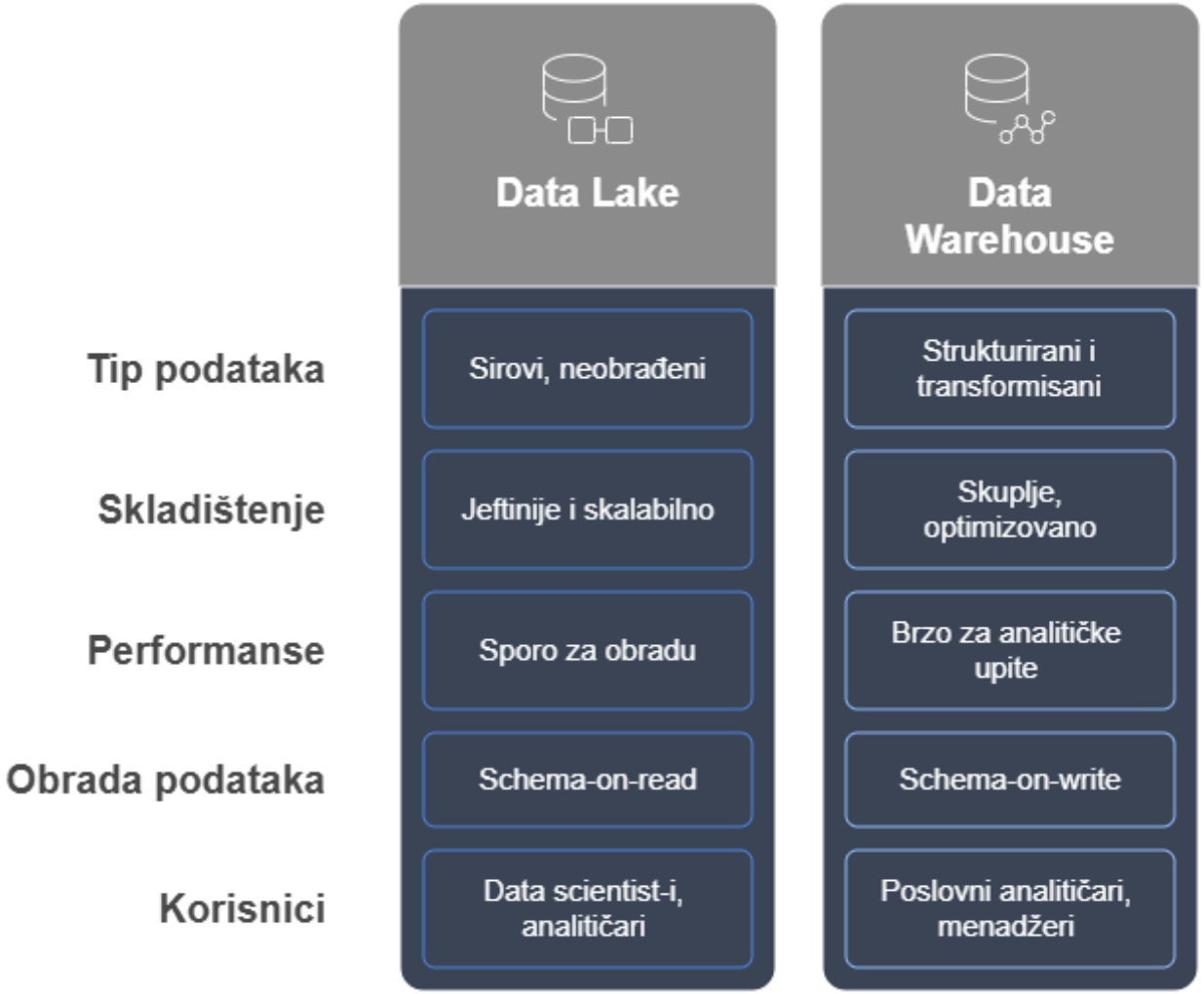


Osobine Data lake sistema

- **Data Lake**
  - Centralizovani repozitorijum za čuvanje struktuiranih i ne struktuiranih podataka
  - Prikupljeni podaci se čuvaju bez promene jer želimo da sačuvamo originalne podatke
    - Tekst, Slike, log podaci, IoT, Video fajlovi, ...
  - Primena je u korišćenju podataka za kreiranje AI i mašinsko učenje
    - E.g. data science – ML\AI
  - Kompanija može da ima warehouse i data lake

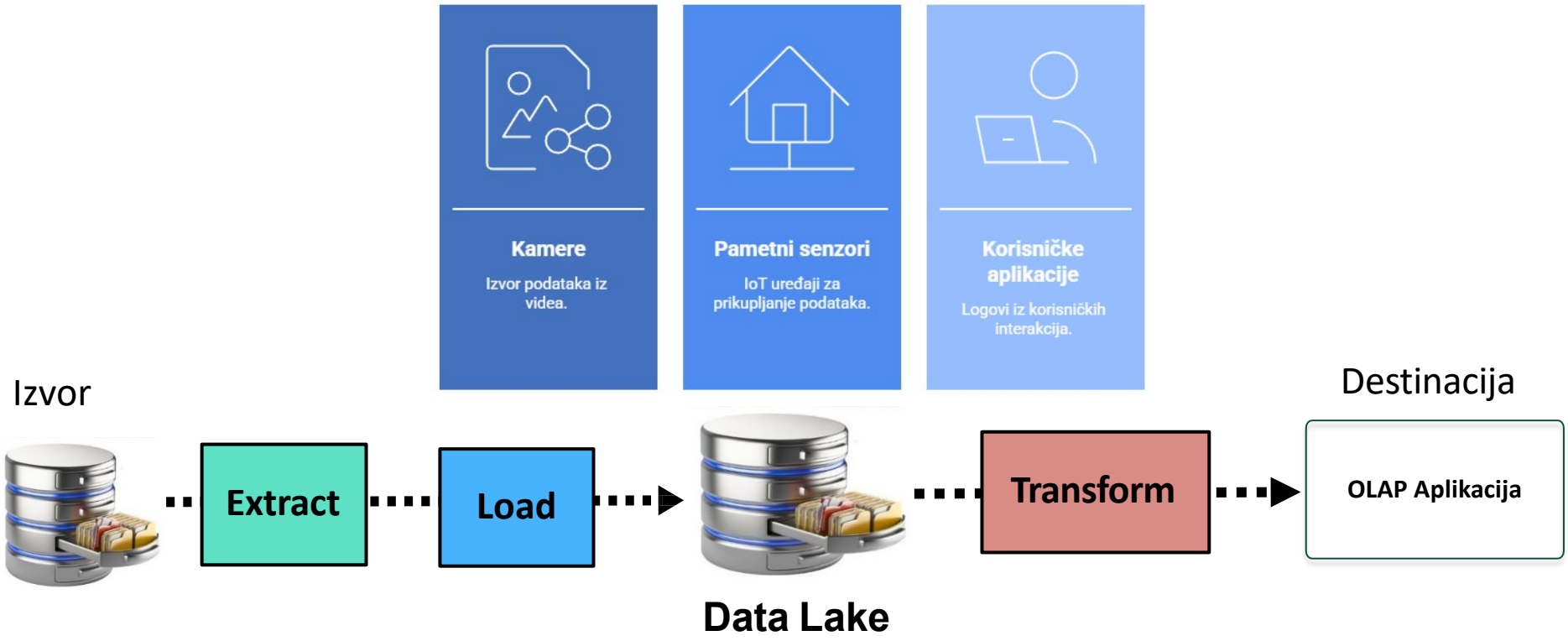


Poređenje Data Lake i Data Warehouse sistema



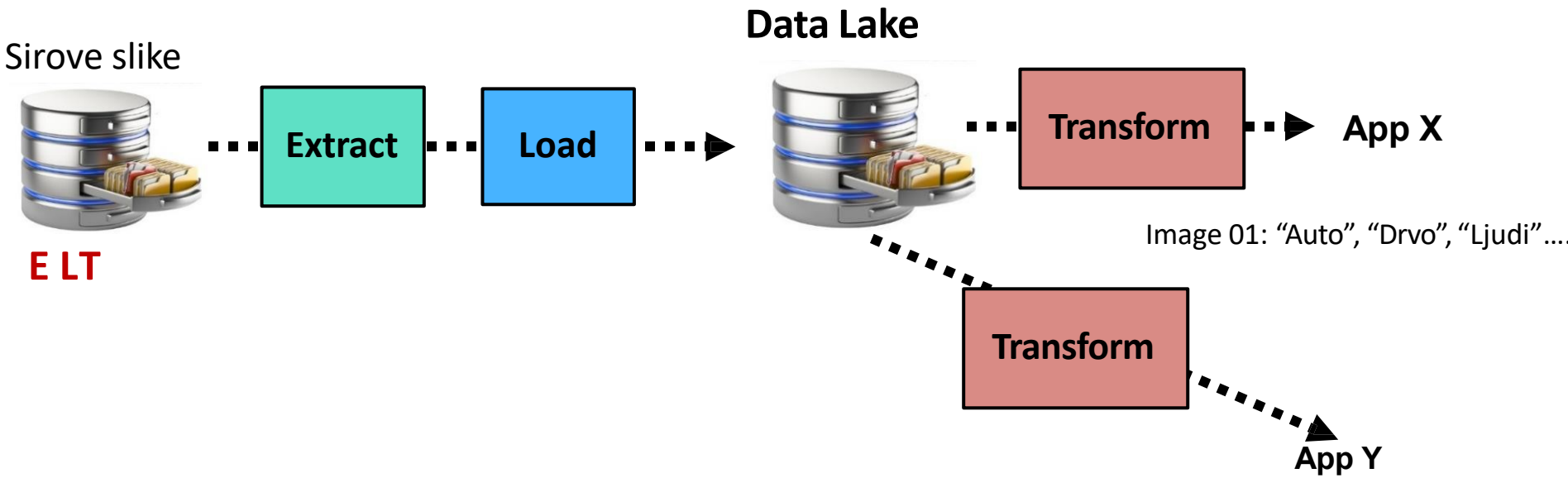
Premeštanje podataka između sistema OLTP - OLAP

- **ELT - Extract, Load, Transform**
  - Proces uzimanja (**extracting**) sirovog podatka (izvor podataka) i unos (**load**) podataka u bazu podataka bez transformacije originalne strukture
  - **Ciljna aplikacija** radi transformaciju strukture podataka



Premeštanje podataka između sistema OLTP - OLAP

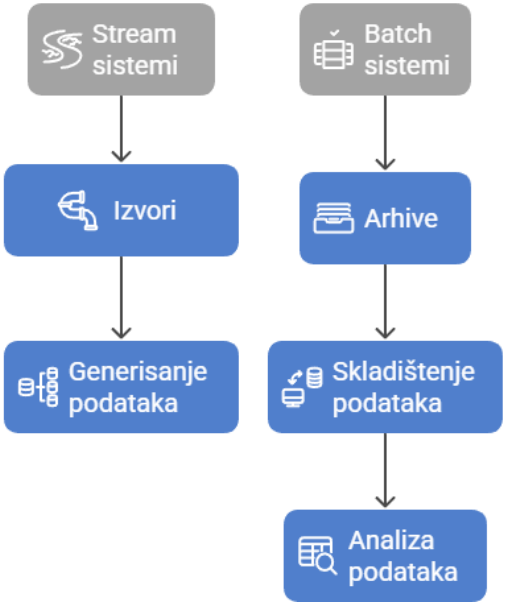
- App X i App Y
  - Postiže se veća fleksibilnost i brža obrada podataka jer je proces transformacije vremenski zahtevan
  - Nedostatak je što se zahteva veći prostor i što podaci nisu optimizovani za datu aplikaciju (use case)



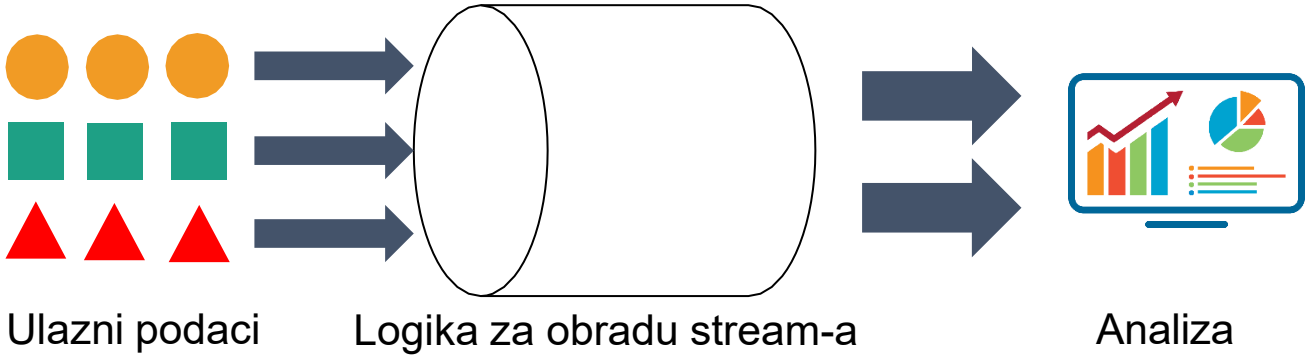
Batch i Stream Obrada

Učestalost premeštanja podataka između sistema

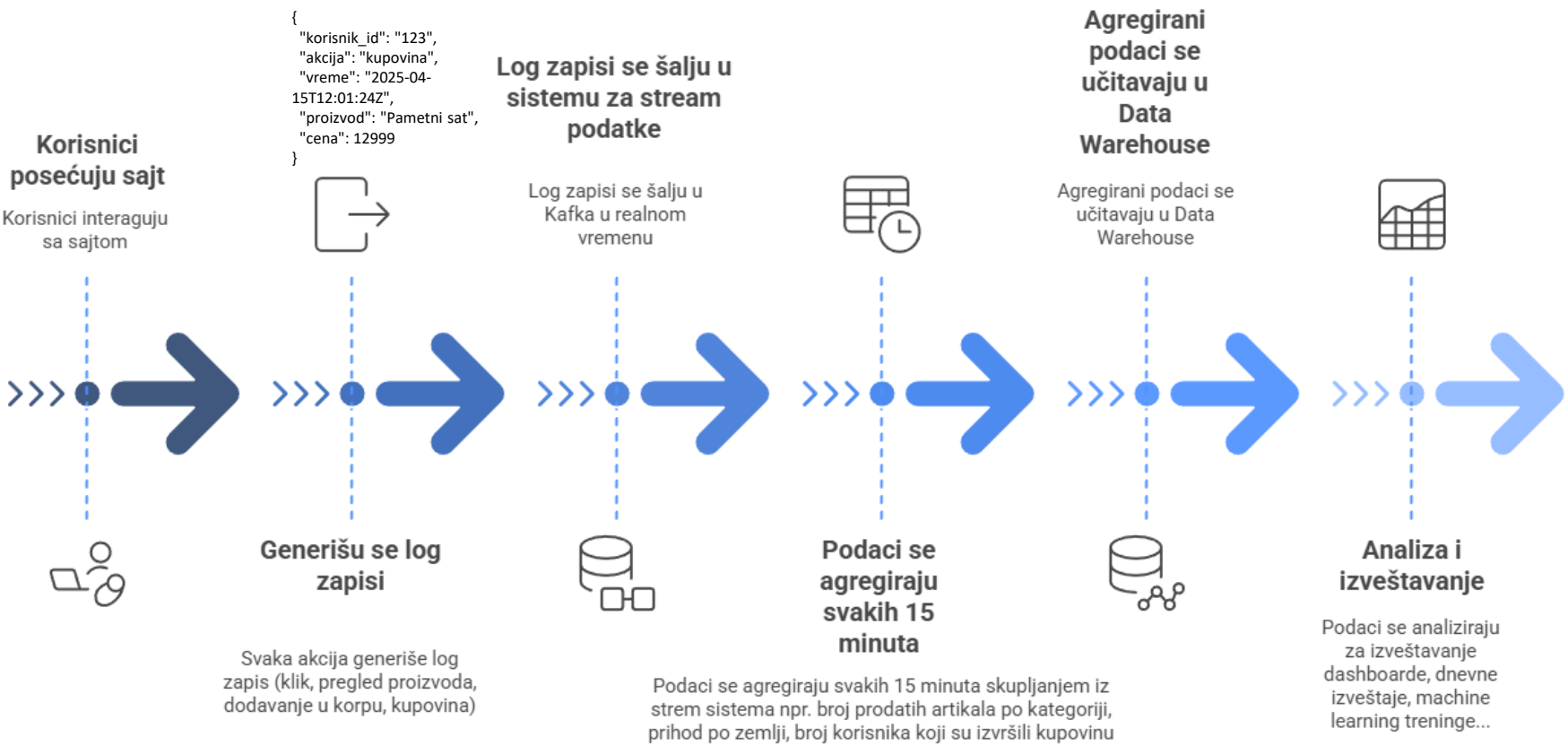
- **Batch obrada**
  - Premeštanje velike količine podataka odjednom na odredišni sistem
  - Obično u noćnim satima (off-peak times), ponavlja se u zakazanim intervalima
    - Npr. na svakih 24h , u 02:00
  - Korisno je ukoliko odredište **ne zahteva podatke u realnom vremenu**
    - Npr. obrada finansijskih podataka u serijama (in batches)



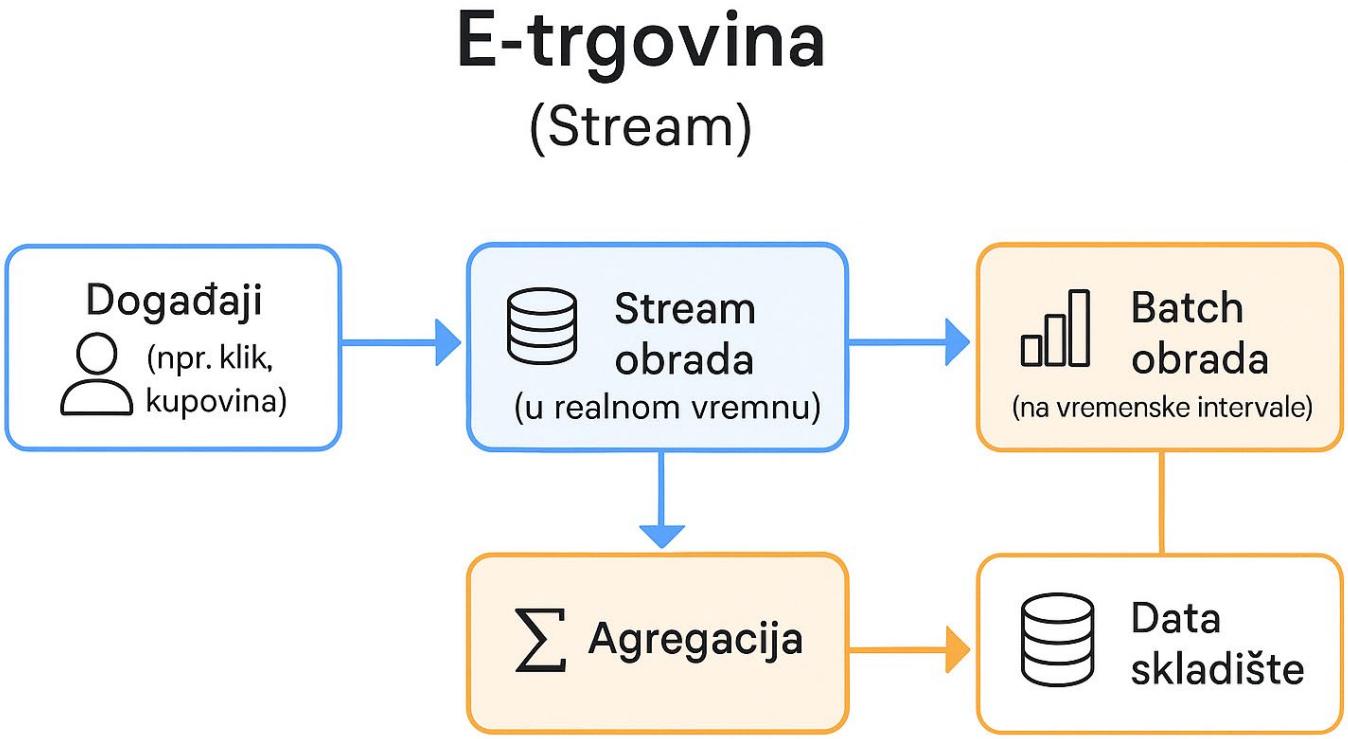
- **Stream obrada**
  - Akcija nad podatkom se sprovodi odmah čim je kreiran
  - Aplikacije u realnom vremenu (**Analytic real time app)**
  - Procesiranje je dizajnirano da se stream konstatno obrađuje



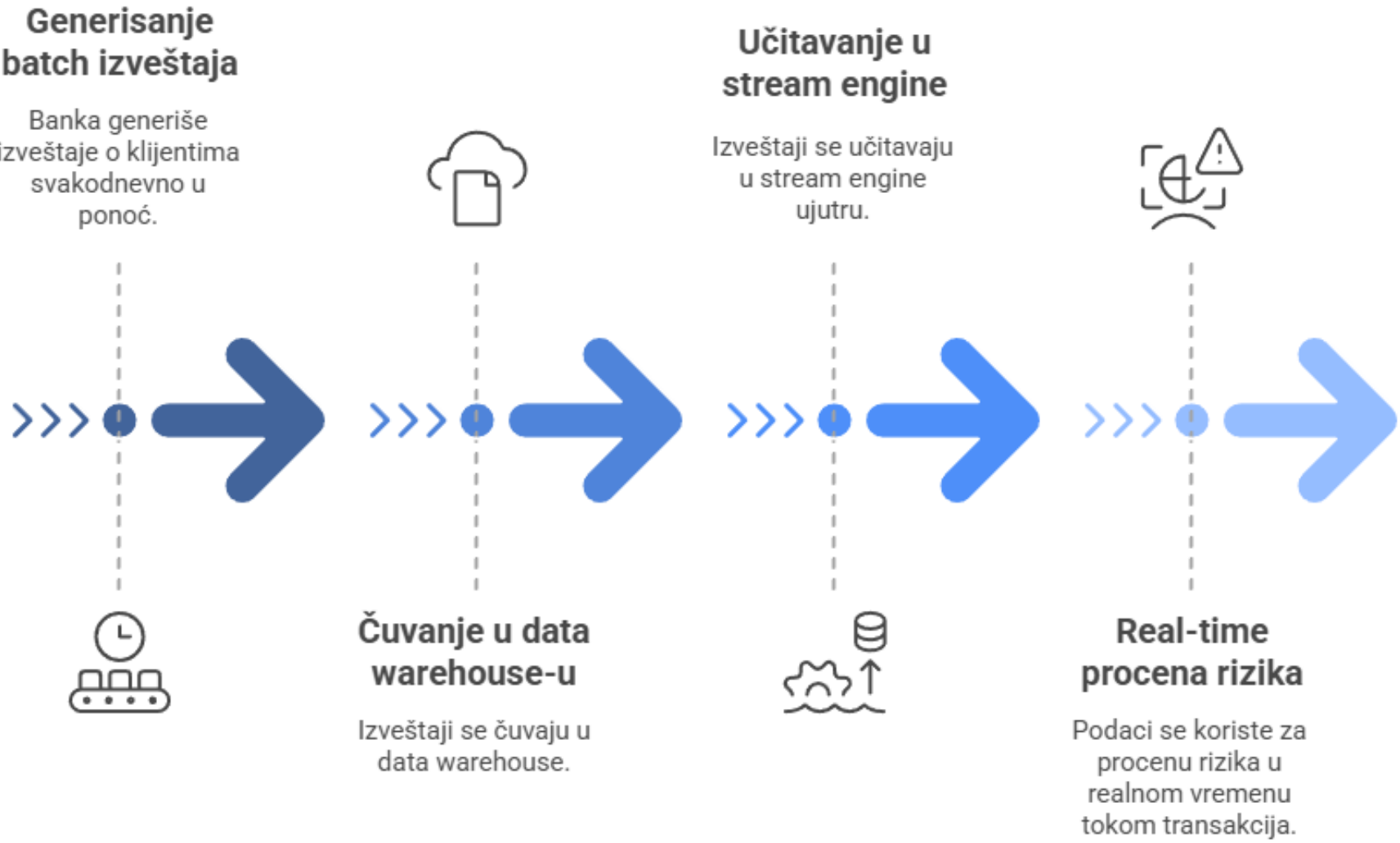
Stream to Batch Obrada



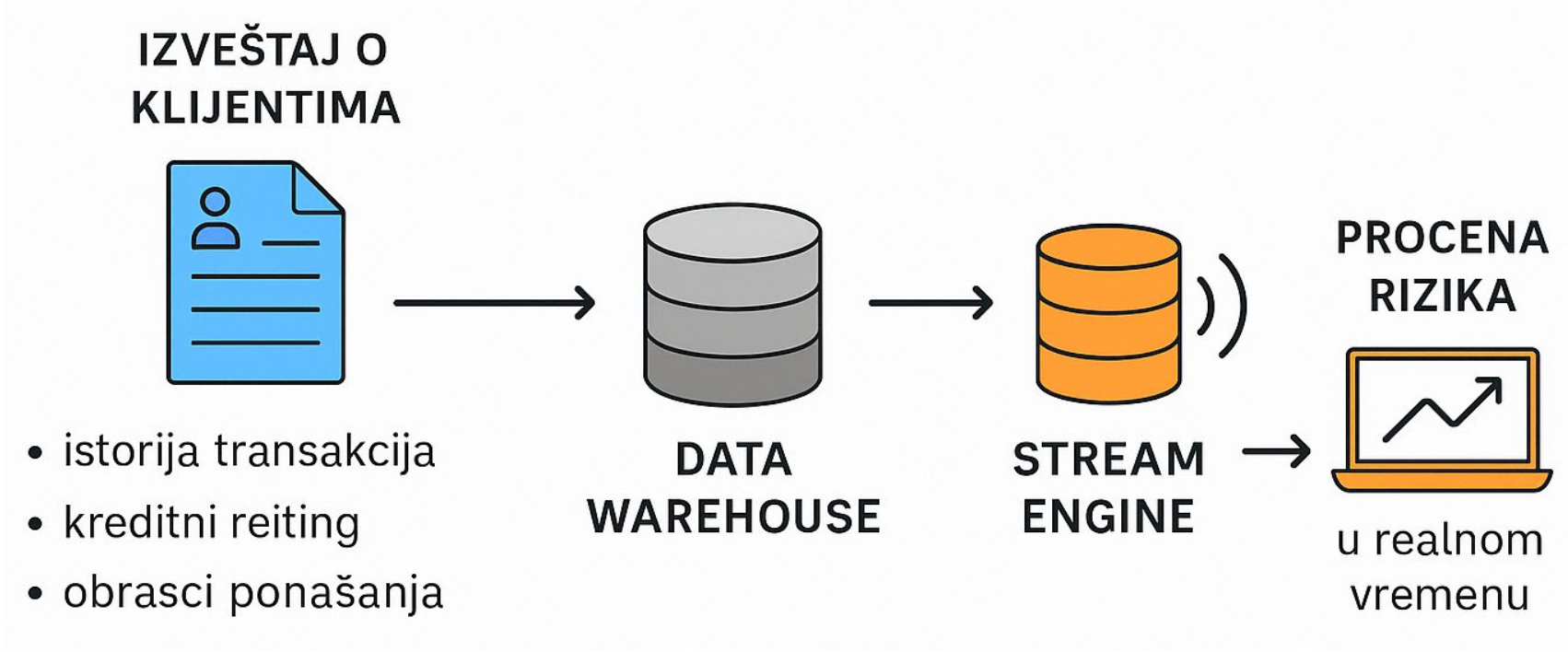
Stream to Batch Obrada



Batch to Stream Obrada

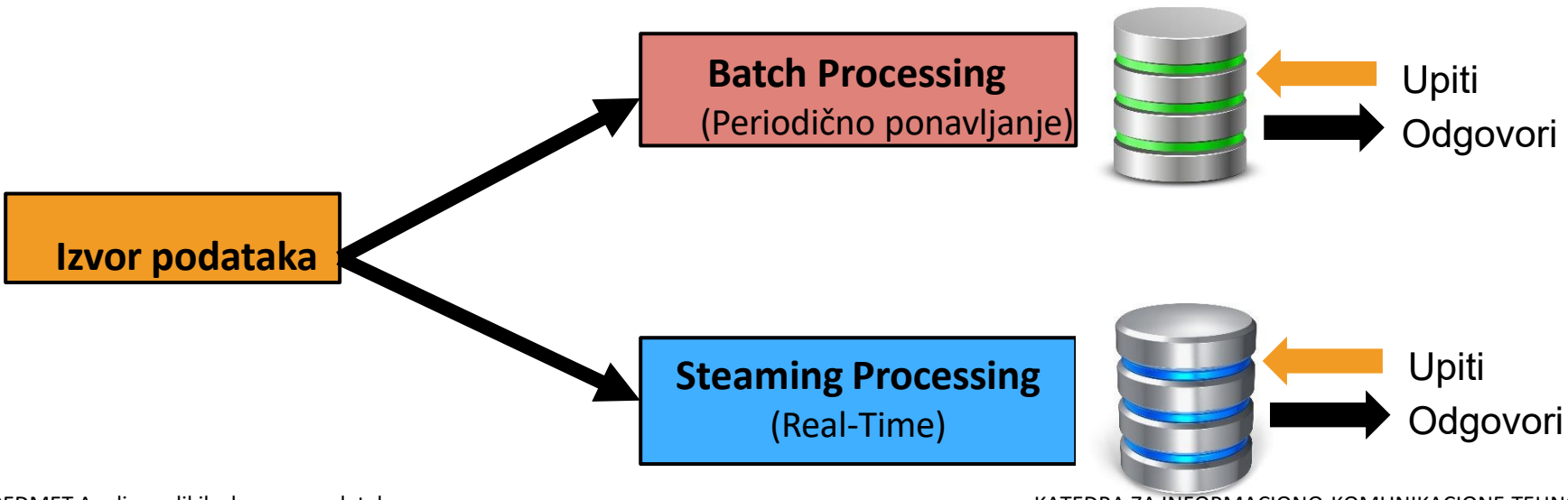


Batch to Stream Obrada



LAMBDA ARHITEKTURA

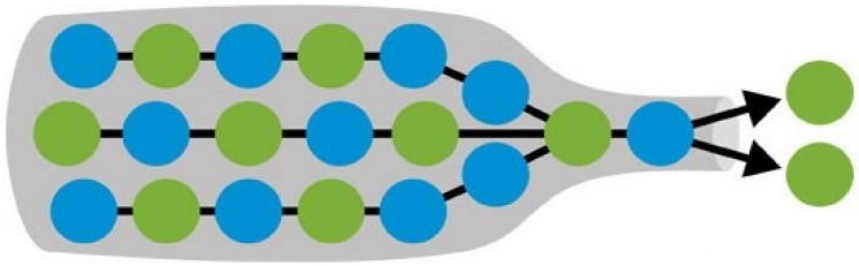
- Sistem kombinuje obe opcije procesiranja paralelno tj obe opcije su ugrađene unutar iste arhitekture
- Arhitektura je dizajnirana da obradi masivnu količinu podataka tako što se koriste prednosti obe metode prenosa(procesuiranja) podataka
- Aplikacije koje rade obradu u realnom vremenu i aplikacije koje radu obradu konsolidovanih podataka
- Koristi se za brzu analizu podataka u realnom vremenu, ali i tačne, kompletne rezultate koji dolaze kroz kasniju, sporiju obradu



SLOJEVI LAMBDA ARHITEKTURE

Skaliranje

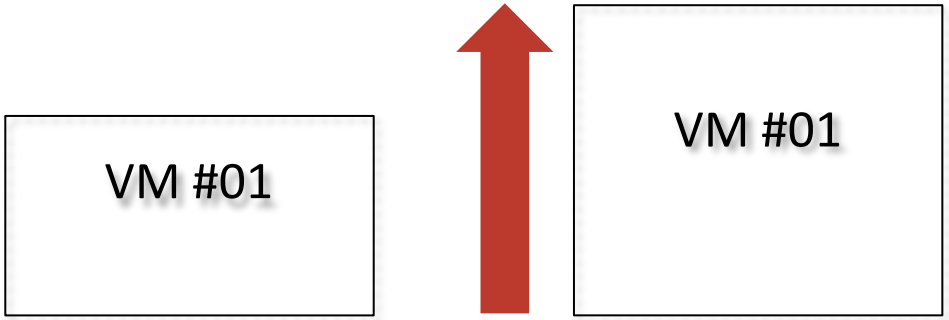
- **Skalabilnost sistema**
  - Skaliranje je proces upravljanja resursima u sistemu da bi se zadovoljile tražene performanse
    - Resursi - CPU, Memorija, Prostor za skladištenje….
    - Prekomerno iskorišćeno (Over-utilized), Nedovoljno iskorišćeno (Under-utilized)
    - Tekući procesi (Ongoing process)
- **Skaliranje Baze podataka**
  - Baze podataka su najčešće **usko grlo** (bottleneck) u aplikaciji
  - Skaliranje baze podataka je vrlo bitno
  - **Dve opcije skaliranja beze podataka**
    - Vertikalno skaliranje
    - Horizontalno skaliranje



Vertikalno Vs Horizontalno Skaliranje

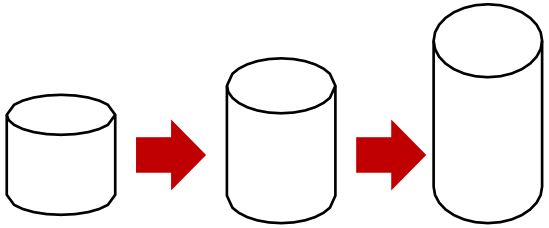
- Skaliranje aplikacije i baze podataka je danas najveći izazov

Vertikalno Skaliranje (up/down)



2 x vCPU
1 x GB RAM

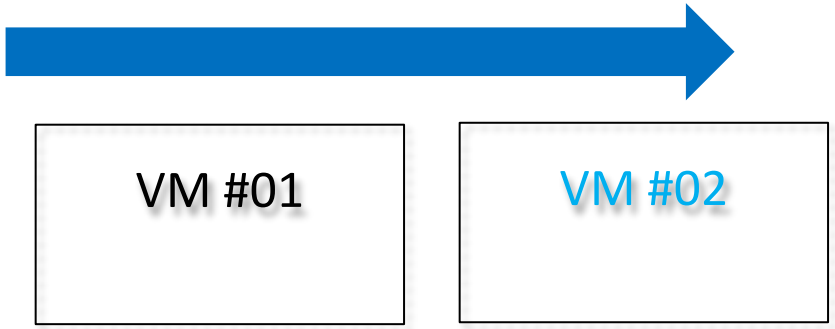
4 x vCPU
2 x GB RAM



2 x vCPU
1 x GB RAM

4 x vCPU
2 x GB RAM

Horizontalno Skaliranje (out/in)



2 x vCPU
1 x GB RAM

2 x vCPU
1 x GB RAM