



BIG DATA – velike količine podataka koji dolaze iz različitih izvora



Sadržaj

- Što je Big Data ?
- Organizacija Big Data procesa
- Standardni koraci i iskustva
- Pitanja

Što je Big Data ?

- Ogromne količine podataka – prevelike ili prekompleksne za obradu tradicionalnim metodama i alatima za obradu podataka
- Ključne karakteristike („3 Vs”)
 1. **Volumen („Volume”)**
 - velike količine podataka (npr. senzori)
 - terabajti, petabajti ili 100M zapisa
 2. **Brzina („Velocity”)**
 - brzina generiranja i obrade podataka
 3. **Raznolikost („Variety”)**
 - različiti formati podataka
 - Strukturirani (RDBMS - tablice)
 - Polustrukturirani (XML, JSON, CSV)
 - Nestrukturirani (slike, videozapisi)

Organizacija Big Data procesa ?

- Standardni koraci:
 1. Definiranje ciljeva i strategije
 2. Prikupljanje podataka
 3. Obrada podataka
 4. Pohrana podataka
 5. Kontrola podataka
 6. Vizualizacija i korištenje
 7. Održavanje i skaliranje
 8. Sigurnost
 9. Kontinuirano poboljšavanje
- Osiguravaju da procesi budu:
 - učinkoviti
 - skalabilni
 - usklađeni s poslovnim ciljevima

Korak 1 - Definiranje ciljeva i strategije

- **Ciljevi**

- PRIMARNI

- kroz **jednu točku** ostvariti pristup **heterogenim** izvorima podataka
 - sučelja:
 - **korisnička** (UI/UX) – grafički/tablični i geografski prikaz
 - **strojna** (API) – REST API

- SEKUNDARNI

- **kontola kvalitete podataka** (numerički podaci sa senzora)
 - **automatska** (Scheduled)
 - **ručna** (On-demand)

- **Strategija**

- **normalizirati** heterogene izvore podataka – što jednostavniji model podataka
 - koristiti **ETL alate** za prikupljanje, transformaciju i pohranu podataka
 - **RDBMS** za pohranu → optimizirani – particioniranje, bez ref. integriteta (na Data tablicama)
 - light-weight web-bazirano rješenje, sa „**dobrim**” **Library-jem** za **iscrtavanje grafova**

Korak 2 – Prikupljanje podataka (1)

- **Izvori podataka**

- numerički podaci (na mikro lokaciji – postaja)
 - senzori
 - opažatelji
- rasterski podaci (određene prostorne rezolucije)
 - snimke većih područja (Radar, Satelit)
 - modeli/analize većih područja (ALADIN, INCA)

- **Vrste podataka**

- strukturirani (RDBMS – PostgreSQL)
- polustrukturirani (XML, CSV, HDF, GRB, GeoTIFF, ...)

Korak 2 – Prikupljanje podataka (2)

- **Sustav u brojkama**

- Senzorika (automatika) / Numerika / **10 min**
 - 500 postaja
 - **10 min podaci:** $500 \times 40 \text{ MME} \times 1 \approx 20.000 \text{ zapisa (2.8 M zapisa/dan)}$
 - **1 sec podaci:** $500 \times 8 \text{ MME} \times 600 \approx 2.400.000 \text{ zapisa (350 M zapisa/dan)}$
- Opažatelji (klasika) / Numerika / **1 dan**
 - 630 postaja
 - **1 dan podaci:** $630 \times 80 \text{ MME} \times 1 \approx 50.000 \text{ zapisa (0,05 M zapisa/dan)}$
 - **1 sat podaci:** $630 \times 10 \text{ MME} \times 24 \approx 150.000 \text{ zapisa (0,15 M zapisa/dan)}$
- Snimke / Rasteri / **5 min**
 - 1 Postaja
 - **5 min podaci:** $1 \times 4 \text{ MME} \times 1 \approx 8 \text{ MB (2,2 GB/dan)}$
- Analitika / Rasteri / **1 sat**
 - 1 Postaja
 - **1 sat podaci:** $1 \times 6 \text{ MME} \times 1 \approx 12 \text{ MB (0,3 GB/dan)}$
- Prognostika / Rasteri / **8 sati**
 - 1 Postaja
 - **1 sat podaci (za 72 sata):** $1 \times 20 \text{ MME} \times 1 \approx 2.880 \text{ MB (2,8 GB/dan)}$


```

1 <?xml version="1.0" encoding="UTF-8"?><Data>
2   <DTDVersion>0001</DTDVersion>
3   <Station>dhmz.bilogora</Station>
4   <DaySet>
5     <Date>20240501</Date>
6     <EqualFormatData>
7       "TOA5","001_1150_karlovac_g","CR1000X","42716","CR1000X.Std.06.00","CPU:main.CRIX","36418","Mjerenja_lmin"
8       "TIMESTAMP","RECORD","datalogger_panel_temperature","precipitation_intensity_RT","precipitation_accu_RT_NRT","precipitation_accu_R"
9       "TS","RN","°C","mm/min","mm","mm","mm","min","","m","","mm/min","°C","%", "min","cm","mm","mm","°C","°C","°C","°C","°C","°C","°C","°C"
10      "","","Smp","Smp","Smp","Smp","Smp","Tot","Tot","Smp","Smp","Smp","Smp","Smp","Smp","Smp","Smp","Smp","Smp","Smp","Smp","Smp","Smp"
11      "2024-05-13 22:13:00",0,21.88,0,0,0,171.46,0.017,1,25850,51,0.001,14.1,96,1,-307,16.7,0,14.09,14,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"
12      "2024-05-13 22:14:00",1,21.88,0,0,0,171.46,0.017,1,27950,51,0.001,14.1,96,0.98,-306.9,16.7,0,14.16,13.97,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"
13      "2024-05-13 22:15:00",2,21.88,0,0,0,171.46,0.017,1,23690,51,0.001,14.1,96,0.89,-306.8,16.7,0,14.12,14.08,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"
14      "2024-05-13 22:16:00",3,21.88,0,0,0,171.46,0.017,1,21870,51,0.001,14.1,96,1,-306.8,16.7,0,14.12,13.82,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"
15      "2024-05-13 22:17:00",4,21.87,0,0,0,171.46,0.017,1,25370,51,0.001,14.1,96,0.77,-306.8,16.7,0,14.18,14.68,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"
16      "2024-05-13 22:18:00",5,21.87,0,0,0,171.46,0.017,1,21790,51,0.001,14.1,96,0.64,-306.8,16.7,0,14.19,14.32,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"
17      "2024-05-13 22:19:00",6,21.87,0,0,0,171.46,0.017,1,20780,51,0.002,14.1,96,1,-306.9,16.8,0.1,14.21,13.75,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"
18      "2024-05-13 22:20:00",7,21.87,0,0,0,171.46,0.017,1,20130,51,0.001,14.2,96,1,-306.9,16.7,-0.1,14.2,14.15,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"
19      "2024-05-13 22:21:00",8,21.87,0,0,0,171.46,0.017,1,22360,52,0.003,14.2,96,1,-307,"NAN","NAN","NAN","NAN",14.51,"NAN","NAN","NAN","NAN","NAN","NAN"
20      "2024-05-13 22:22:00",9,21.88,0,0,0,171.46,0.017,1,15370,61,0.006,14.2,96,1,-307.1,"NAN","NAN","NAN","NAN",13.95,"NAN","NAN","NAN","NAN","NAN","NAN"
21      "2024-05-13 22:23:00",10,21.89,0,0,0,171.46,0.017,1,15840,61,0.007,14.2,96,1,-307,16.8,0,14.28,13.77,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"
22      "2024-05-13 22:24:00",11,21.9,0,0,0,171.46,0.017,1,14850,61,0.007,14.1,96,1,-307.1,16.7,-0.1,14.15,14.05,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"
23      "2024-05-13 22:25:00",12,21.91,0,0,0,171.46,0.017,1,13150,61,0.004,14.1,96,1,-307.2,"NAN","NAN","NAN","NAN",13.47,"NAN","NAN","NAN","NAN","NAN","NAN"
24      "2024-05-13 22:26:00",13,21.92,0,0,0,171.46,0.017,1,13690,61,0.005,14.1,96,1,-307,16.7,0,14.18,14.11,"NAN","NAN","NAN","NAN","NAN","NAN","NAN"

```

Korak 3 – Obrada podataka (1)

- **ETL – Extract Transform Load**
- Koraci



1. Definicija procesa transformacije

- ulazni formati podataka (CSV, XML, RDBMS, GeoTIFF, HDF, GRB ...)
- postupak transformacije (numeričke / prostorne)
- učestalost/frekvencija izvršavanja (5 min, 10 min,...)
- željeni izlazni format podataka (za pohranu)

2. Mapiranje ulaznih podataka u odnosu na izlazne

- **vrijednosti iz tablice/kolone u CSV-u na ulazu** se moraju učitati točno na **određenu tablicu/kolonu na izlazu**
- **Kompleksno mapiranje ! ☹**

Korak 3 – Obrada podataka (2)

3. Odabir alata

- gotovo ETL rješenje

- **PROS:**

- Vizualni prikaz procesa
- Puno gotovih komponenti
- Brzi rezultati osnovnih funkcionalnosti

- **CONS:**

- Kompleksni procesi !!??
- Nedovoljno „fino“ upravljanje resursima !!!

- vlastito rješenje → **Programski jezik (+ Library)**

- **PROS:**

- „Fino“ upravljanje resursima
- **Vectorized processing ! 😊**

- **CONS**

- Pisanje vlastitog koda – nema vizualnog prikaza procesa 😞

Korak 3 – Obrada podataka (3)

4. Konfiguracije

- specifikacija po kojoj rade alati → veza između ulaza i izlaza
- sadrže sve potrebne informacije za izvršavanje ETL procesa

Korak 4 – Pohrana podataka (1)

- **Arhitektura – Zvezdasta schema**

- uobičajena u dizajnu skladišta podataka

- centar – činjenica (fact) - mjerenje

- atributi:

- termin

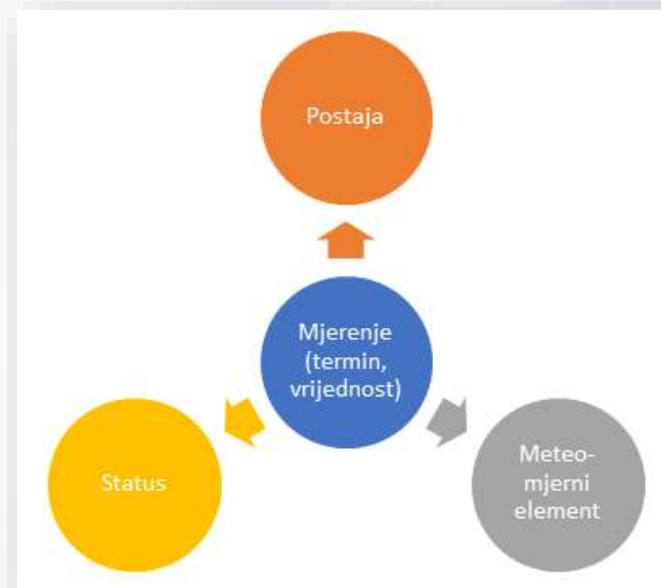
- vrijednost

- dimenzije

- meteo-mjerni element (MME)

- postaja

- status



- **Tehnologija za pohranu**

- PostgreSQL (RDBMS)



Korak 4 – Pohrana podataka (2)

- **Organizacija / optimizacija PostgreSQL**

- Data tablice

- Vrijednosti mjerenja → **✗ REF. INTEGRITET**
 - **PARTICIONIRANJE** → ovisno o frekvenciji uzorkovanja podataka
 - **AGREGIRANJE** → na senzorima / u Sustavu
 - **RETENCIJA** → brisanje podataka starijih od određenog perioda
 - **BRZI / SPORI DISKOVI**

ataka

DHMZ **Centralna integracijska platforma**

PRIVATE

METEO PODACI

MODUL

Boy

→ [Collins](#)

 DHMZ Centralna integracijska platforma

POČETNA

PRIMA 7

METEO PODACI

7.10

Ad

negative effect.



Ručna kontrola kvalitete

Peunați în odații N

PresentJant

Statistika I validacija ▼



szwerni 5446

Postaja: BILDOGRA

classificacja DHMZ

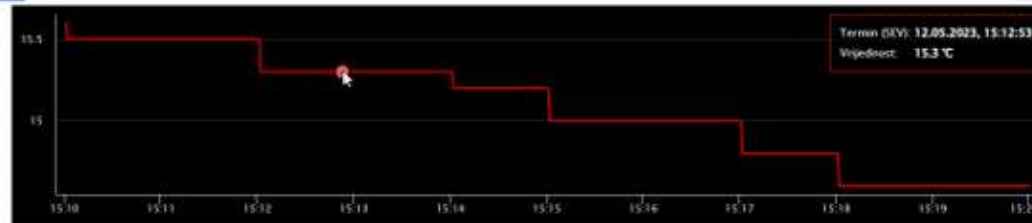
izvor	AMP - sažetak (27)
-------	--------------------

NAME	Totmeska vrijednost temperature zraka na 2 m
UNIT	°C/m

Tip no.	TR
---------	----

intervju objavljuje. Sekundni

Month, year	°C
1998	19.9
1999	19.9
2000	19.9
2001	19.9
2002	19.9
2003	19.9
2004	19.9
2005	19.9
2006	19.9
2007	19.9
2008	19.9
2009	19.9
2010	19.9
2011	19.9
2012	19.9
2013	19.9
2014	19.9
2015	19.9
2016	19.9
2017	19.9
2018	19.9
2019	19.9
2020	19.9
2021	19.9
2022	19.9
2023	19.9
2024	19.9
2025	19.9
2026	19.9
2027	19.9
2028	19.9
2029	19.9
2030	19.9
2031	19.9
2032	19.9
2033	19.9
2034	19.9
2035	19.9
2036	19.9
2037	19.9
2038	19.9
2039	19.9
2040	19.9
2041	19.9
2042	19.9
2043	19.9
2044	19.9
2045	19.9
2046	19.9
2047	19.9
2048	19.9
2049	19.9
2050	19.9
2051	19.9
2052	19.9
2053	19.9
2054	19.9
2055	19.9
2056	19.9
2057	19.9
2058	19.9
2059	19.9
2060	19.9
2061	19.9
2062	19.9
2063	19.9
2064	19.9
2065	19.9
2066	19.9
2067	19.9
2068	19.9
2069	19.9
2070	19.9
2071	19.9
2072	19.9
2073	19.9
2074	19.9
2075	19.9
2076	19.9
2077	19.9
2078	19.9
2079	19.9
2080	19.9
2081	19.9
2082	19.9
2083	19.9
2084	19.9
2085	19.9
2086	19.9
2087	19.9
2088	19.9
2089	19.9
2090	19.9
2091	19.9
2092	19.9
2093	19.9
2094	19.9
2095	19.9
2096	19.9
2097	19.9
2098	19.9
2099	19.9
2100	19.9



Termin (SV): 12.05.2023, 15:12:53
Vrijednost: 15.3 °C

[Dohvati podatke](#)



Tablični/grafički prikaz meteo-mjernih elemenata > Numerički modeli

DHMZ CIP

1.0.0 GAS

meteorološki podaci

Dohvat meteoroloških podataka

GET

/mjerenja/numericka Numerička mjerenja prema lokaciji i datumu

GET

/mjerenja/rasteri Rasterski prikaz određenih meteoroloških elemenata

GET

/mjerenja/rasteri/dostupni Prikaz dostupnih rasterskih datoteka

GET

/mjerenja/izvorni-podaci Izvorni meteorološki podaci

GET

/mjerenja/izvorni-podaci/dostupni Prikaz dostupnih rasterskih datoteka

preuzimanje podataka

Servisi za kreiranje i dohvat .zip datoteka s željenim podacima

POST

/preuzimanja-podataka Kreiraj novi zahtjev za preuzimanjem

GET

/preuzimanja-podataka Dohvati dostupne zahtjeve za preuzimanje

GET

/preuzimanja-podataka/{preuzimanjeId} Dohvati zahtjev za preuzimanjem

GET

/preuzimanje-podataka/{preuzimanjeId}/datoteka Dohvati datoteku s podacima

GET

/mjerenja/statusi Dostupni statusi mjerenja

GET

/mjerenja/statusi/{statusId} Dostupni statusi mjerenja

Korak 7 – Održavanje i skaliranje

- **Neophodan** redoviti monitoring i maintenance baze (preventivno + korektivno) te ostalih komponenti
- VM okruženje omogućuje proširenje resursa
- Load balancing arhitektura komponenti sustava

Korak 8 – Sigurnost

- **Sigurnosni model → autorizacija nad mjerenjima**
 - mogućnost **fine granulacije** prava pristupa podacima s obzirom na:
 - **Postaju** – npr. podaci nekih investitorskih postaja nisu za generalnu javnost
 - **Meteo-mjerni element** – npr. elementi kao status releja su od koristi samo internim korisnicima za administraciju/održavanje
 - **Status** – npr. neka mjerenja su testna ili neslužbena te ne smiju biti vidljiva javnosti

Korak 9 – Kontinuirano poboljšavanje

- **Evaluacija**

- redovito procjenjivanje performansi sustava i pronalazak „uskih grla”

- **Optimizacija**

- provođenje optimizacija s obzirom na povratne informacije od korisnika

Pitanja ?

