

# Tema 1

## Analiza exploratorie a datelor pentru oferte imobiliare Mineritul datelor și analiza datelor (MDAD)

2023 - v1.0

Deadline: 30.03.2023 (23:59)

## Descriere generală

Pentru această tema va trebui să realizați o analiză exploratorie de date pentru informații legate de oferte imobiliare. Rezolvarea corectă și completă a temei presupune implementarea următorilor pași:

1. Colectarea datelor prin conectarea și interacțiunea cu un serviciu de date pus la dispoziție de noi
2. Transformarea datelor
  - a. Din format JSON în formatul necesar pentru restul pipeline-ului
  - b. Descoperirea și corectarea erorilor care au apărut din procedura de colectare
  - c. Adăugarea de noi coloane (acolo unde este cazul, prin transformarea celor existente)
3. Stocarea datelor într-un sistem/format ales de voi
4. Analiza datelor obținute
5. Prezentarea analizei realizate într-un raport tehnic

## Colectarea datelor

Infrastructura pentru livrarea datelor este formată din 2 api-uri care se găsesc pe mașina cu adresa IP 141.85.224.171. Puteți vedea metodele disponibile pentru fiecare la:

- <http://141.85.224.171:8000/docs>
- <http://141.85.224.171:8001/docs>

Pentru obținerea datelor va trebui întâi să obțineți cheia de autentificare de la serviciul disponibil la adresa <http://141.85.224.171:8000/key>. Se va trimite un **POST request** cu adresa de e-mail cu care sunteți înrolat pe Moodle iar în cazul în care aceasta este validă veți primi cheia sub forma unui string de tip *UUID4*. Structura mesajelor o regăsiți în exemplul următor:

**Request:** {"identifier" <string>: "email@stud.acs.upb.ro"}

**Response:** {"status\_code": 200, "key": <UUID-4> "0161ec73-b223..."}

```
Curl
curl -X 'POST' \
  'http://141.85.224.171:8000/key' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
    "identifier": "dan_teodor.ponc@upb.ro"
  }'
```

Request URL

http://141.85.224.171:8000/key

Server response

Code	Details
200	<p>Response body</p> <pre>{   "key": "5f5d7e22-0116-4717-b103-43e48bbe71c" }</pre> <p>Response headers</p> <pre>content-length: 46 content-type: application/json date: Mon, 20 Mar 2023 17:43:03 GMT server: uvicorn</pre>

Apoi veți folosi această cheie la adresa <http://141.85.224.171:8001/dmda/a1/data> pentru a descărca datele:

[illegible]

```
Request: {
  "identity": {
    "Identifier_value" <string>: "email@stud.acs.upb.ro",
    "Key" <UUID-4>: "0161ec73-b223..."
  }
}
```

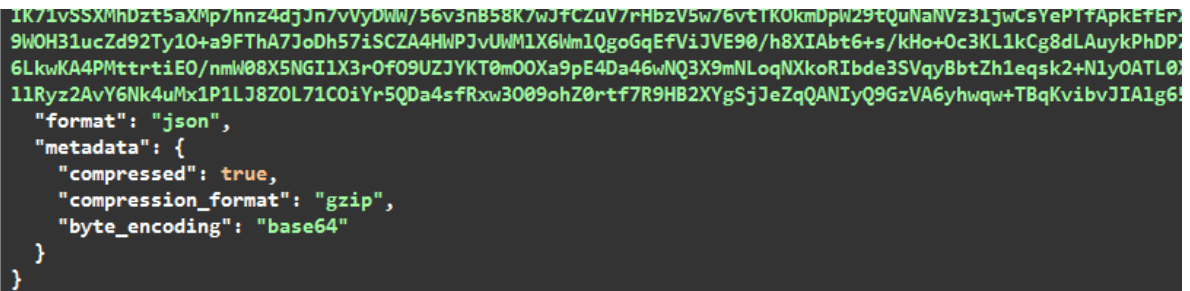
```
Response: {
  status_code: 200
  "data": "H4sIAIO+aGGQC/+y9a3PiyLYt+lcU/...",
```

```

"format": "json",
"metadata": {
  "compressed": true,
  "compression_format": "gzip",
  "byte_encoding": "base64"
}
}

```

Datele sunt transmise sub formă de bytes în formatul specificat de “byte\_encoding”. Acești bytes pot sau nu să fie compriși, conform detaliilor din metadata. Câmpul format indică ce tip de obiect a fost serializat:



```

IK71vSSXMHdZt5aXmp/hnz4djJn7vVyDWW/56v3nB58K/wJfCZuV/rHbzV5w/6vE1KOkmDpW29tQuNaNVz3IjwCsYeP1fApkEter...
9WOH31ucZd92Ty10+a9FTThA7Jodh57iSCZA4HWPJvUWM1X6Wm1QgoGqEFViJVE90/h8XIAbt6+s/kHo+0c3KL1kCg8dLAuykPhDP...
6LkwKA4PMttrtiE0/nmW08X5NGI1X3r0f09UZJYKT0m00Xa9pE4Da46wNQ3X9mNLoqNXkoRIbde3SVqyBbtZh1eqsk2+N1y0ATL0...
11Ryz2AvY6Nk4uMx1P1LJ8ZOL71COiYr5QDa4sfRkw3009ohZ0rtf7R9HB2XYgSjJeZqQANIyQ9GzVA6yhwqw+TBqKvibvJIA1g6...
  "format": "json",
  "metadata": {
    "compressed": true,
    "compression_format": "gzip",
    "byte_encoding": "base64"
  }
}

```

## Transformarea datelor

Datele cu care veți interacționa conțin diverse tipuri de erori (pe care va trebui să le descoperiți voi). De exemplu, prețul extras pentru oferte (care eventual poate fi utilizat ca variabilă țintă într-un set de date construit pe baza acestor date) este pentru unele exemple raportat ca preț metru pătrat iar pentru altele ca preț pentru întregul imobil. O astfel de eroare se poate remedia parțial, de obicei, prin înlocuirea prețului ofertelor sub 20k € cu produsul dintre preț și suprafața utilă a imobilului. Unul din obiectivele acestei teme este să descoperiți problemele din setul de date și să le corectați prin metode pe care va trebui să le explicați în raportul tehnic.

## Analiza datelor

În analiza datelor sunteți liberi să vă inspirați din ceea ce am discutat la curs și la prezentările practice, însă NU VĂ LIMITAȚI strict la procedurile de acolo. Sunteți liberi (chiar încurajați) să explorați și alte idei legate de tratamentul și analiza datelor pe care le aveți la dispoziție.

# Raport tehnic

Pe lângă fișierele cu codul sursă al temei voastre, va trebui să predați și un raport tehnic care să conțină rezultatele analizei făcute de voi. Raportul tehnic va fi realizat fie sub forma unui fișier de tip PDF conținând toate graficele, tabelele și informațiile din analiza voastră fie sub forma unui jupyter notebook (scris în Python) similar cu cele prezentate la curs. În acest raport tehnic veți descrie, pe lângă rezultatele obținute, procedurile pe care le-ați urmat pentru a colecta, transforma, stoca și analiza datele. Pe Moodle veți încărca o arhivă *.zip* cu numele vostru și prefixul *Assignment \_1\_* (e.g. *Assignment\_1\_Ionel\_Popescu.zip*) în care veți include codul sursă și raportul tehnic.