Web scraping, cloud storage and analysis

Assignment 1 Deliverables

Part 1
Data collection

○ Perform **Web Scraping** techniques on 100 entries
○ Create a **data frame** suitable for scraped data
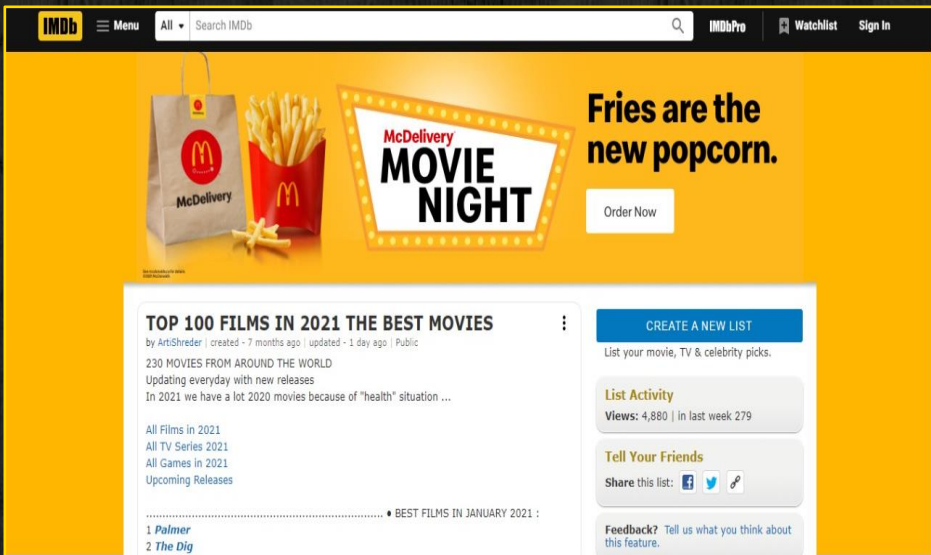○ Platform of choice – **Beautiful Soup**

Part 2
Data Storage

○ Create **database** for using the data frame
○ Platform of choice - **MongoDB**

Part 3
Web Application Development

○ Connecting the database to a **UI**
○ Performing **Create, Read, Update & Delete** functions
○ Platform of choice - **Flask**

Data Source

Source Data – *Best 100 movies of 2021 by IMDb*

Technologies & Libraries Used

Technologies
- Python
- Flask
- Bootstrap
- MongoDB Atlas

Libraries
- Beautiful Soup
- PyMongo
- Flask
- pandas

1. Data Collection

Web Scraping
Data Frame

HTML Inspection and tags finding

- Film Title
- IMDB Rating
- Run Time
- Genre
- Short Narration
- Director

Data labels

Web Scraping
Platform used – Beautiful Soup

○ Source code used for web scraping

```
web_scraping.py                    ×    untitled

 1   from urllib.request import urlopen as uReq
 2   from bs4 import BeautifulSoup as soup
 3   import pymongo
 4   from pymongo import MongoClient
 5   import pandas as pd
 6
 7
 8   my_url = 'https://www.imdb.com/list/ls084462066/'
 9
10   #opening up connection, grabbing the page
11
12   uClient = uReq(my_url)
13
14   page_html = uClient.read()
15
16   uClient.close()
17
18   #html parser
19
20   page_soup = soup(page_html, "html.parser")
21
22   # grabs each product
23
24   fab = page_soup.findAll("div",{"class":"lister-item mode-detail"})
25
26   #exporting to CSV
27
28   filename = "products.csv"
29
30   f = open(filename, "w")
31
32   headers = "Film Title,IMDB Rating,Run Time,Genre,Short Narration,Director\n"
33
34   f.write(headers)
35
36
```

7

Source code used for web scraping

```
37
38
39   for fa in fab:
40
41       try:
42           title = fa.div.img["alt"]
43
44           rating = fa.findAll("span", {"class":"ipl-rating-star__rating"})[0].text
45
46           runtime = fa.findAll("span", {"class":"runtime"})[0].text
47
48           genre = fa.findAll("span", {"class":"genre"})[0].text.strip()
49
50           description = fa.findAll("p", {"class":""})[0].text.strip()
51
52           director = fa.findAll("p", {"class":"text-muted text-small"})[1].a.text
53
54
55
56           print ( " Film Title : " + title)
57           print (" IMDB Rating : " + rating)
58           print (" Run Time : " + runtime)
59           print (" Genre : " + genre)
60           print (" Short Narration : " + description)
61           print (" Director : " + director)
62
63           f.write(title.replace(",", " ") + "," + rating + "," + runtime + "," + genre.replace(",", "|")  + "," + description.replace
64
65
66       except IndexError:
67           print ("_")
68
69
70   f.close()
71
72
```

Web Scraping

.csv file created after scraping data

Data Frame

○ Importing .csv to data frame & sorting the data to MongoDB

```
61          print (" Director : " + directer)
62
63          f.write(title.replace(",", " ") + "," + rating + "," + runtime + "," + genre.replace(",", "|")  + "," + description.replace
64
65
66      except IndexError:
67          print ("_")
68
69
70  f.close()
71
72
73
74  # Import CSV Files into Pandas Dataframes
75
76
77  df = pd.read_csv ('products.csv', encoding = "ISO-8859-1")
78
79  df
80
81
82  # connecting to mongo DB and storing the data
83
84  cluster = MongoClient("mongodb+srv://testuser:testuser123@cluster0.dj0nr.mongodb.net/myFirstDatabase?retryWrites=true&w=majority")
85
86  db  = cluster["testDb"]
87
88  collection = db["testData"]
89
90  collection.insert_many(df.to_dict('records'))
91
```

Web Scarping

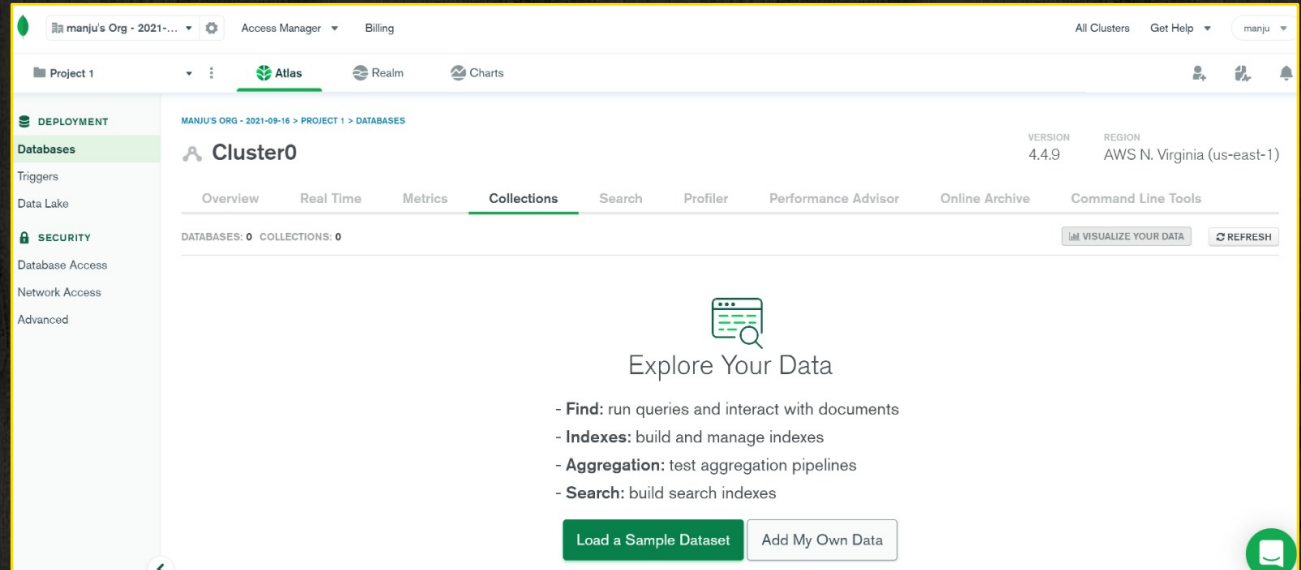## Data frame created using the .csv file

2. Data Storage

Database Creation

Database Creation

Platform used - MongoDB
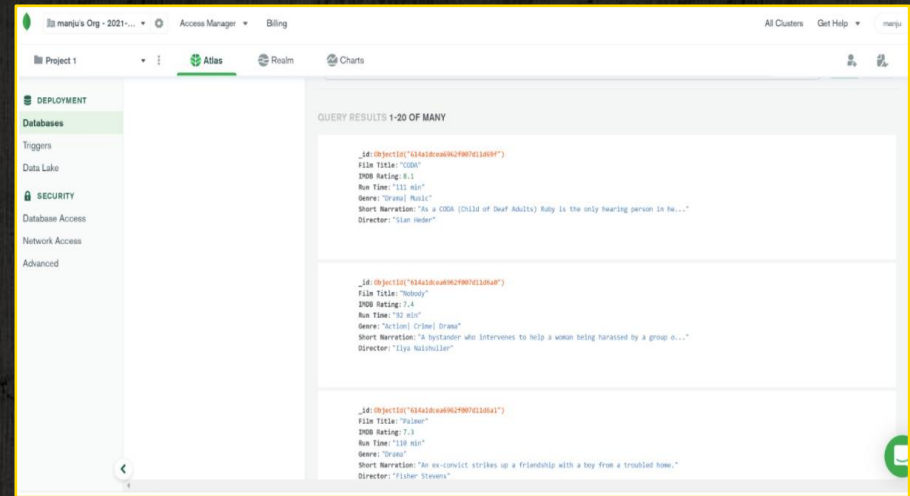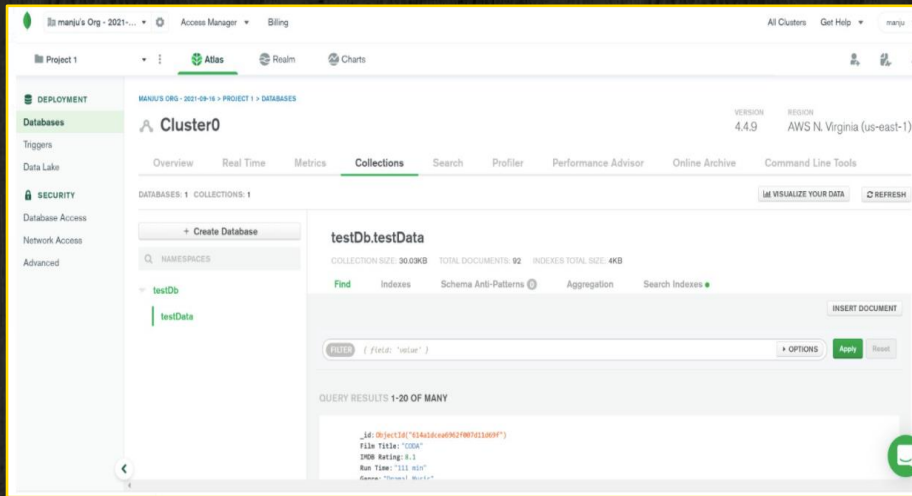
○ MongoDB Atlas – before web scraping

Database Creation

○ MongoDB Atlas – after loading web scraped data

3. Web Application Development

User Interface

Web App Development

User Interface

o View – view all available information available on the database

o Add – add new records into the database

o Edit – make changes to existing records on the database

o Delete – remove existing records from the database

Web App Development

Platform used - Flask

○ Downloading .csv file into IDE

Web App Development

UI - View

○ View Page

IMDB Top 100 films in 2021 - The Best Movies

Add Movie

Film Title	IMDB Rating	Run Time	Genre	Short Narration	Director		
CODA	8.1	111 min	Drama| Music	As a CODA (Child of Deaf Adults) Ruby is the only hearing person in her deaf family. When the family's fishing business is threatened Ruby finds herself torn between pursuing her love of music and her fear of abandoning her parents.	Sian Heder	Edit	Delete
Nobody	7.4	92 min	Action| Crime| Drama	A bystander who intervenes to help a woman being harassed by a group of men becomes the target of a vengeful drug lord.	Ilya Naishuller	Edit	Delete
Palmer	7.3	110 min	Drama	An ex-convict strikes up a friendship with a boy from a troubled home.	Fisher Stevens	Edit	Delete
Seaspiracy	8.2	89 min	Documentary| Adventure	Passionate about ocean life a filmmaker sets out to document the harm that humans do to marine species - and uncovers alarming global corruption.	Ali Tabrizi	Edit	Delete
Love Death & Robots	8.5	15 min	Animation| Short| Action	A collection of animated short stories that span various genres including science fiction fantasy horror and comedy.	Scott Whyte	Edit	Delete
Halston	7.5	240 min	Biography| Drama	The story of a man who leverages his single invented name into a worldwide fashion empire that's synonymous with luxury sex status and fame literally defining the era.	Ewan McGregor	Edit	Delete
Time	8.3	59 min	Crime| Drama	Eric is a prison officer who tries to protect those in his charge. When one of the most dangerous inmates identifies his weakness Eric faces an impossible choice between his principles and his love for his family.	Sean Bean	Edit	Delete
Joji	7.9	113 min	Crime| Drama	Joji an engineering dropout and the youngest son of a rich family lives with his aspirations of becoming super wealthy. Driven by greed and blind ambition he decides to execute his plans following an unexpected event	Dileesh Pothan	Edit	Delete

Web App Development

○ view.html Code

○ View Flask Code

Web App Development

UI - Add

○ Add Page

Please enter the below details to add a movie

Film Title:

IMDB Rating:

Run Time:

Genre:

Short Narration:

Director:

Add movie

Web App Development

○ Entering data to Add Page

○ New Movie Updated on Add Page

Please enter the below details to add a movie

Film Title:

new movie

IMDB Rating:

8

Run Time:

10 min

Genre:

action

Short Narration:

test narration

Director:

manju

Add movie

the Titans		min	Adventure	fight the Arcane Order for control over the magic that binds them all.			Delete
Resident Alien	8.2	44 min	Comedy| Drama| Mystery	A crash-landed alien named Harry who takes on the identity of a small-town Colorado doctor and slowly begins to wrestle with the moral dilemma of his secret mission on Earth.	Alan Tudyk	Edit	Delete
Don't Breathe 2	6.1	98 min	Horror| Thriller	The sequel is set in the years following the initial deadly home invasion where Norman Nordstrom (Stephen Lang) lives in quiet solace until his past sins catch up to him.	Rodo Sayagues	Edit	Delete
Holler	6.2	90 min	Drama	To pay for her education and the chance of a better life a young woman joins a dangerous scrap metal crew.	Nicole Riegel	Edit	Delete
The Conjuring: The Devil Made Me Do It	6.3	112 min	Horror| Mystery| Thriller	The Warrens investigate a murder that may be linked to a demonic possession.	Michael Chaves	Edit	Delete
Them	7.3	45 min	Drama| Horror| Thriller	A Black family moves to an all-white Los Angeles neighborhood where malevolent forces next door and otherworldly threaten to taunt ravage and destroy them.	Deborah Ayorinde	Edit	Delete
A Dark Foe	6.7	114 min	Crime| Drama| Horror	A guilt-ridden FBI agent stranded in the painful memory of the abduction of his sister suffers from a rare condition known as Nyctophobia an irrational fear of the dark and will have to... See full summary »	Maria Gabriela Cardenas	Edit	Delete
Akilla's Escape	5.1	90 min	Crime| Drama	In a crime-noir about the urban child-soldier Akilla Brown captures a fifteen-year-old Jamaican boy in the aftermath of an armed robbery. Over one gruelling night Akilla confronts a cycle of generational violence he thought he escaped.	Charles Officer	Edit	Delete
Annette	6.4	141 min	Drama| Musical| Romance	A stand-up comedian and his opera singer wife have a 2 year old daughter with a surprising gift.	Leos Carax	Edit	Delete
The Colony	5.4	104 min	Sci-Fi| Thriller	Set in the distant future a female astronaut shipwrecked on the long-decimated Earth must decide the fate of the wasteland's remaining populace.	Tim Fehlbaum	Edit	Delete
new movie	8	10 min	action	test narration	manju	Edit	Delete

21

Web App Development

○ New Movie Updated on MongoDB
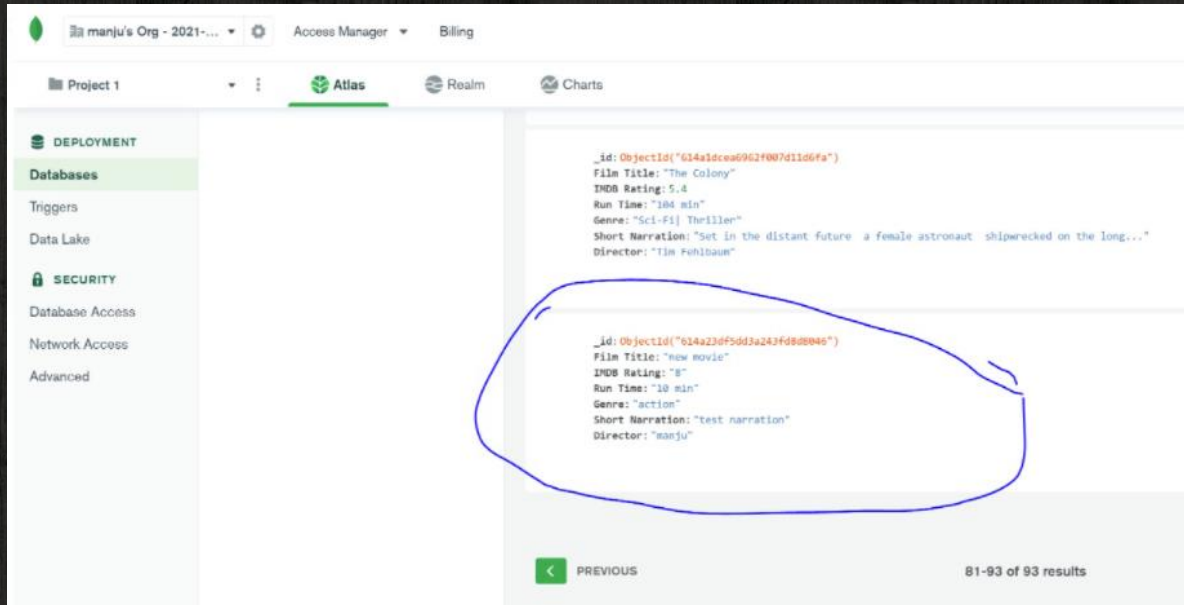
Web App Development

○ add.html Code

○ Add Flask Code

Web App Development

UI - Edit

○ Edit Page

Edit the movie details

Film Title	IMDB Rating	Run Time	Genre	Short Narration	Director	
CODA	8.1	111 min	Drama| Music	As a CODA (Child of [	Sian Heder	Save

Web App Development

○ Edit Movie Name to CODA edit

○ CODA edit Updated on Edit Page

Web App Development

o   CODA edit Updated on MongoDB

Web App Development

○ edit.html Code   ○ Edit Flask Code

Web App Development

Function - Delete

○ Page Before Deleting   ○ DB Before Deleting

Web App Development

○ Page After Deleting

○ DB After Deleting

Web App Development

○ delete.html Code

○ Delete Flask Code

```
            </form>
        </td>
        <td>
            <form action="{{url_for('delete_movie', id=movie._id)}}" method="post">
                <input type="hidden" name="_method" value="DELETE">
                <input type="submit" value="Delete" class="btn btn-danger">
            </form>
        </td>
    </tr>
    {% endfor %}
</table>
{% endblock %}
```

```
33    # Delete Movie
34    @app.route('/delete_movie/<string:id>', methods=['POST','GET'])
35  ∨ def delete_movie(id):
36        print(id)
37        myquery = { "_id": ObjectId(id) }
38        print(myquery)
39        collection.delete_one(myquery)
40        return redirect(url_for('index'))
41
```

References

https://www.google.com/search?q=imdb+logo&rlz=1C1CHBF_enCA972CA972&source=lnms&tbm=isch&sa=X&ved=2ahUKEwieksyU0JDzAhW7GVkFHS_fDVIQ_AUoAXoECAEQAw&biw=1280&bih=632&dpr=1.5

https://www.slidescarnival.com/prospero-free-presentation-template/91

Thank You!