DBMS : CS3563 : Assignment 2 Report

GROUP 4

Name	Roll no.
Raj Patil	CS18BTECH11039
Vedant Singh	CS18BTECH11047
Karan Bhukar	CS18BTECH11021
Himanshu Bishnoi	CS16BTECH11018

Describing Deliverables

1. *DBMS A2 Report Grp4.pdf*
2. *create.sql*
    - *Creating the raw data tables and the main schema tables*
3. *importing_raw.txt*
    - *Importing data from the preprocessed tsvs into the raw tables*
4. *main.sql*
    - *populating the main tables*
5. *fkey.sql*
    - *Altering the main tables to enforce foreign key constraints*
    - *One also needs to validate this constraint explicitly which in turn triggers the foreign key validation for the data.*
6. *get_movie_info.py*
    - *Collecting websites and plot info for some Generic_Media entries*

Creating Empty Database

Before importing the data, we created all the raw tables and the ones corresponding to our ER diagram submitted in the last assignment. This step also involves adding validation constraints for only primary keys. We add the foreign key constraints after populating the tables due to the reasons mentioned below:

- There were some media-ID's listed as TV Series-ID's in the title_episodes table but these were not present in the original 'title_basics' table from where all the information about TV Series has been taken from. So, we need to backtrack and add these extra keys back into our Generic Media table so that there can exist a valid foreign key constraint.

EXAMPLE:

Here we highlight the sequence of enforcing the constraint with respect to various events for a single table(Generic Media)

```
BEGIN;

CREATE TABLE public."Generic_Media"
(
    "IMDB_id" character varying NOT NULL,
    rating real,
    original_title character varying,
    antisocial_elements character varying[],
    languages character varying[],
    genres character varying[],
    plot_outline character varying,
    "PG_rating" character varying,
    runtime bigint,
    PRIMARY KEY ("IMDB_id")
);
< ..... Creating other tables ..... >

< ..... Populating data from the raw data tables .....>



ALTER TABLE public."Generic_Media_Location"
    ADD FOREIGN KEY ("Generic_Media_IMDB_id")
    REFERENCES public."Generic_Media" ("IMDB_id")
    NOT VALID;
```

```
ALTER TABLE public."Movie"
    ADD FOREIGN KEY (movie_id)
    REFERENCES public."Generic_Media" ("IMDB_id")
    NOT VALID;

ALTER TABLE public."Person_Generic_Media"
    ADD FOREIGN KEY ("Generic_Media_IMDB_id")
    REFERENCES public."Generic_Media" ("IMDB_id")
    NOT VALID;

ALTER TABLE public."TV_Series"
    ADD FOREIGN KEY (series_id)
    REFERENCES public."Generic_Media" ("IMDB_id")
    NOT VALID;
< ..... Altering other tables .....>

END;
```

Cleaning and Importing into raw data tables

Summary:

1. Cleaning data, handling specific cases to make them parsable
2. Importing them via PgAdmin4, filling the raw data tables (not a part of the schema)

Specifics:

a. **multivalued attributes : preprocessing involved enclosing them in braces**
b. **lines containing empty strings: preprocessing involved replacing them with the conventional “\N”**
c. **UTF-8 encoding was used with the “|” (vertical pipe) being the quote and escape character; all other special characters(#,$,@ and so on) were already used in the data provided and we couldn’t use quotes (“) as the data also contained instances of multiple double quotes**

EXAMPLE :
Importing data into raw data tables

```
--command " "\\copy public.name_basics (nconst, \"primaryName\",
\"birthYear\", \"deathYear\", \"primaryProfession\", \"knownForTitles\")
FROM <path to data> DELIMITER E'\\t' CSV HEADER ENCODING 'UTF8' QUOTE '|'
NULL '\\N' ESCAPE '|';""
```

Modifications in the original ERD

The way in which the data was present in the raw files inspired us to make some small changes in our original ERD.

- Originally, we decided to have two different tables corresponding to TV_Series_Location and Movie_Location, but we decided to merge this into a single relation which related Generic Media and Location.
- Since it was originally given that runtime would be different for different regions, we had the attribute of runtime in the relation which related Generic Media and Location. But seeing the data provided to us, we decided to add this attribute to the entity itself.
- Rather than adding a new id for TV series and movies and then adding a foreign key to Generic Media, we decided to make the same key as primary and foreign key for the TV_Series and Movies tables.
- We have also added two extra attributes in the Episodes table (Season no. and episode no.) for quicker querying.
- Rather than creating a separate table for websites, we have included them as a multivalued attribute in Generic_Media table. The reason being that websites don't have an explicit unique ID associated with them (realised when scraping for extra data) and enforcing uniqueness constraints on a separate websites entity would be difficult.

Populating the Main Relation Sets

Rather than importing data into our tables directly from the given '.tsv' files, we first create and populate the tables corresponding to the given 7 files.

We then use SQL queries to populate the data from these tables into the tables created from our ERD. Query wise explanation is provided in the history.sql file.

EXAMPLE:
Importing the data of episodes from the table 'title_basics'

```
INSERT INTO public."Episodes" (episode_id, original_title, runtime)
SELECT tconst, "originalTitle", "runtimeMinutes"
FROM public."title_basics" AS B
WHERE B."titleType" = 'tvEpisode';
```

Now that we have the basic data of each episode, we access other tables to get more relevant data such as the ID of the TV series this episode belongs to, or the rating of this episode.

```
UPDATE
    public."Episodes"
SET
    series_id = R."tconstParent"
FROM (
    SELECT
       tconst, "tconstParent"
    FROM
       public."Episodes" AS E
       INNER JOIN
       public."title_episodes" AS R
       ON E.episode_id = R.tconst
    ) AS R
WHERE episode_id = R.tconst;
```

Scraping extra data

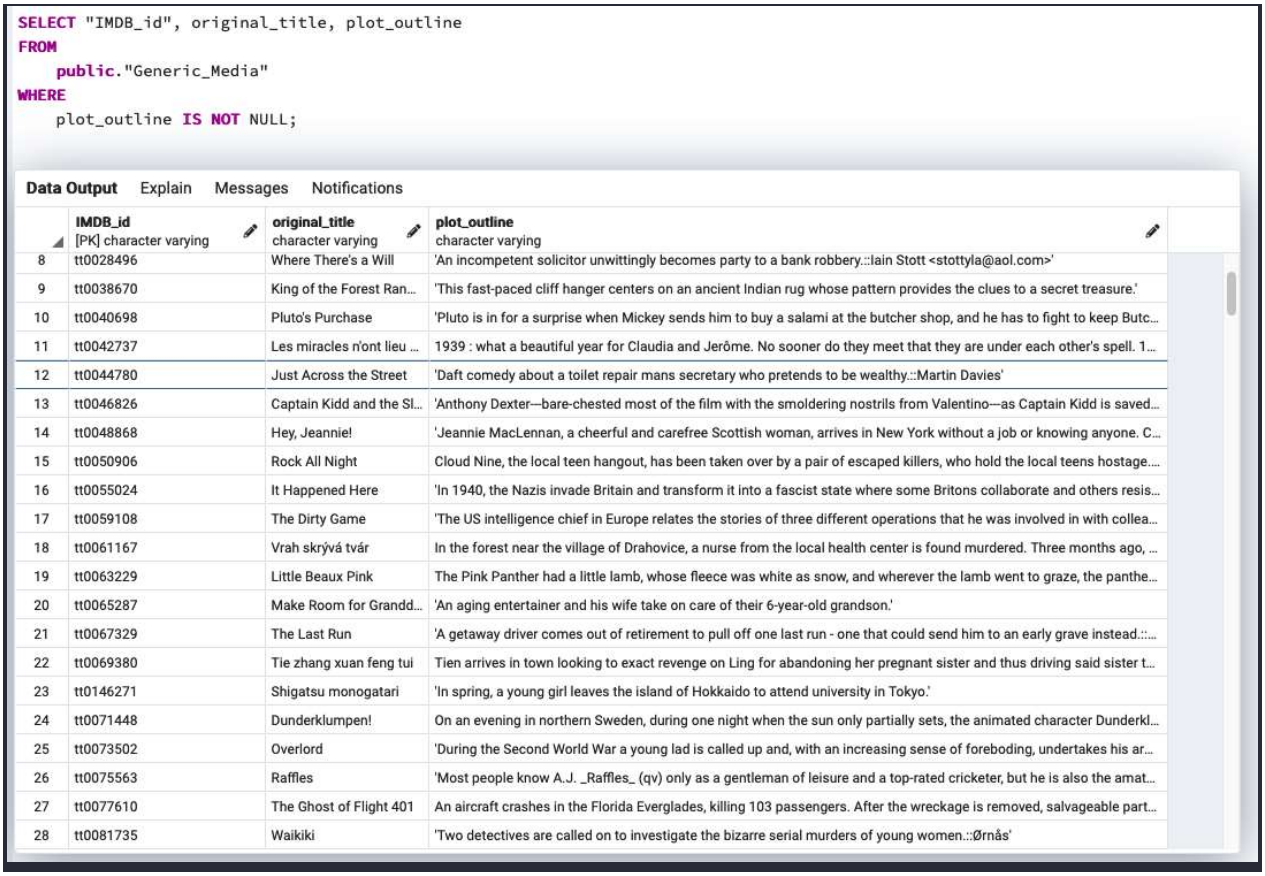
Logistics :

- We were able to locate data for the entities Generic_Media, Person, Websites and Company.(via imdbpy)
- Note that we could not find a single preprocessed tsv containing the eccentric missing data in the relation sets.
- Therefore, we had to dispatch an http request for a particular detail for a single entity each time.
- There are around 7 million media ids in this database. Each request was associated with a response time of 3 seconds (using asynchronous IO and writing in a decoupled manner) i.e. a total of 243 days.
- That is not feasible. So we have sampled for every 2000th entry upto 4 million entries for the sake of pedagogical purposes.
- We will be submitting a script for scraping websites and plots from the API.

EXAMPLE :
Updating websites and plots of generic media and similarly for public.”Episodes”

```
UPDATE public."Generic_Media" GM
SET plot_outline = R.plot, websites = R.websites
FROM
    (
    SELECT M."IMDB_id", INF.plot, INF.websites
    FROM
       public."Generic_Media" M
       INNER JOIN
       public.extra_info INF
       ON M."IMDB_id" = INF."IMDB_id"
    ) AS R
WHERE
    GM."IMDB_id" = R."IMDB_id";
```

```sql
SELECT "IMDB_id", original_title, plot_outline
FROM
    public."Generic_Media"
WHERE
    plot_outline IS NOT NULL;
```



```sql
SELECT "IMDB_id", original_title, plot_outline
FROM
    public."Generic_Media"
WHERE
    plot_outline IS NOT NULL;
```

Data Output | Explain | Messages | Notifications

	IMDB_id [PK] character varying	original_title character varying	plot_outline character varying
8	tt0028496	Where There's a Will	'An incompetent solicitor unwittingly becomes party to a bank robbery.::Iain Stott <stottyla@aol.com>'
9	tt0038670	King of the Forest Ran...	'This fast-paced cliff hanger centers on an ancient Indian rug whose pattern provides the clues to a secret treasure.'
10	tt0040698	Pluto's Purchase	'Pluto is in for a surprise when Mickey sends him to buy a salami at the butcher shop, and he has to fight to keep Butc...
11	tt0042737	Les miracles n'ont lieu ...	1939 : what a beautiful year for Claudia and Jerôme. No sooner do they meet that they are under each other's spell. 1...
12	tt0044780	Just Across the Street	'Daft comedy about a toilet repair mans secretary who pretends to be wealthy.::Martin Davies'
13	tt0046826	Captain Kidd and the Sl...	'Anthony Dexter—bare-chested most of the film with the smoldering nostrils from Valentino—as Captain Kidd is saved...
14	tt0048868	Hey, Jeannie!	'Jeannie MacLennan, a cheerful and carefree Scottish woman, arrives in New York without a job or knowing anyone. C...
15	tt0050906	Rock All Night	Cloud Nine, the local teen hangout, has been taken over by a pair of escaped killers, who hold the local teens hostage....
16	tt0055024	It Happened Here	'In 1940, the Nazis invade Britain and transform it into a fascist state where some Britons collaborate and others resis...
17	tt0059108	The Dirty Game	'The US intelligence chief in Europe relates the stories of three different operations that he was involved in with collea...
18	tt0061167	Vrah skrývá tvár	In the forest near the village of Drahovice, a nurse from the local health center is found murdered. Three months ago, ...
19	tt0063229	Little Beaux Pink	The Pink Panther had a little lamb, whose fleece was white as snow, and wherever the lamb went to graze, the panthe...
20	tt0065287	Make Room for Grandd...	'An aging entertainer and his wife take on care of their 6-year-old grandson.'
21	tt0067329	The Last Run	'A getaway driver comes out of retirement to pull off one last run - one that could send him to an early grave instead.::...
22	tt0069380	Tie zhang xuan feng tui	Tien arrives in town looking to exact revenge on Ling for abandoning her pregnant sister and thus driving said sister t...
23	tt0146271	Shigatsu monogatari	'In spring, a young girl leaves the island of Hokkaido to attend university in Tokyo.'
24	tt0071448	Dunderklumpen!	On an evening in northern Sweden, during one night when the sun only partially sets, the animated character Dunderkl...
25	tt0073502	Overlord	'During the Second World War a young lad is called up and, with an increasing sense of foreboding, undertakes his ar...
26	tt0075563	Raffles	'Most people know A.J. _Raffles_ (qv) only as a gentleman of leisure and a top-rated cricketer, but he is also the amat...
27	tt0077610	The Ghost of Flight 401	An aircraft crashes in the Florida Everglades, killing 103 passengers. After the wreckage is removed, salvageable part...
28	tt0081735	Waikiki	'Two detectives are called on to investigate the bizarre serial murders of young women.::Ørnås'

Main Tables (top 10 rows)

Generic Media:

Data Output | Explain | Messages | Notifications

	IMDB_id [PK] character vary	rating real	original_title character varying	antisocial_elem character varying	languages character varying[	genres character varying[]	plot_outline character varying	PG_ratinc character	runtime bigint	websites character varyi
1	tt0000001	5.6	Carmencita	[null]	[null]	{Documentary,Short}	'Performing on what lo...	[null]	1	[null]
2	tt0016241		Pretty Ladies	[null]	[null]	{Comedy,Drama}	Maggie, a headlining c...	[null]	74	[null]
3	tt0018258	5.3	En perfekt gentleman	[null]	[null]	{Drama}	'The old Luny Palace in...	[null]	105	[null]
4	tt0022355	6.1	The Secret Call	[null]	[null]	{Drama}	City Commissioner Fra...	[null]	70	[null]
5	tt0026455	5.9	Harmony Lane	[null]	[null]	{Adventure,Biography,M...	'The life and loves of c...	[null]	89	[null]
6	tt0028496	6.6	Where There's a Will	[null]	[null]	{Comedy}	'An incompetent solicit...	[null]	80	[null]
7	tt0038670	7.1	King of the Forest Ran...	[null]	[null]	{Adventure}	'This fast-paced cliff h...	[null]	167	[null]
8	tt0040698	6.5	Pluto's Purchase	[null]	[null]	{Animation,Comedy,Fam...	'Pluto is in for a surpris...	[null]	7	[null]
9	tt0042737	6.6	Les miracles n'ont lieu ...	[null]	[null]	{Drama,Romance,War}	1939 : what a beautiful...	[null]	98	[null]
10	tt0044780	8	Just Across the Street	[null]	[null]	{Comedy}	'Daft comedy about a t...	[null]	79	[null]

Editable column

Episodes:

Data Output | Explain | Messages | Notifications

	episode_id [PK] character vary	rating real	plot_outline character varying	PG_rating character va	original_title character varying	series_id character vary	runtime bigint	season_number integer	episode_number integer
1	tt0657507	8.2	'A gang of female priso...	[null]	Hard Cell	tt0115285	45	5	19
2	tt0858288	6.6	Lupin searches for the ...	[null]	Hunt Down the Counte...	tt0159175	24	1	10
3	tt0959715	6.7	Duncan is feeling stres...	[null]	Waring Goes Private	tt0101083	[null]	1	7
4	tt0568389	8.4	Quinn investigates the ...	[null]	Every Dog Has Its Day	tt0115161	[null]	1	5
5	tt0505307	8.1	'Malloy and Reed take ...	[null]	Log 144: Bank Robbery	tt0062539	30	2	24
6	tt0513491	6.7	A old female peddler is...	[null]	The Waco Kid	tt0046578	30	3	22
7	tt0684038	7.8	'Spinelli takes the plac...	[null]	Don't Ask Me	tt0126170	[null]	4	31
8	tt0734673	7.4	'A U.S. army lieutenant ...	[null]	The Purple Testament	tt0052520	25	1	19
9	tt0781910	7.7	'The second unaired pil...	[null]	My Fair Munster: Unair...	tt0057773	[null]	[null]	[null]
10	tt0649428	8.5	'Rogers displays an as...	[null]	118	tt0062588	[null]	1	118

Locations:

Data Output Explain Messag

	country [PK] character varying
1	RU
2	KP
3	DK
4	CSXX
5	SN
6	SI
7	CZ
8	KR
9	BS
10	VE

Movies:

Data Output Explain Messages Notifications

	movie_id [PK] character varying	box_office_collection money	budget money
1	tt3392132	[null]	[null]
2	tt3392136	[null]	[null]
3	tt3392138	[null]	[null]
4	tt3392142	[null]	[null]
5	tt3392146	[null]	[null]
6	tt3392166	[null]	[null]
7	tt3392174	[null]	[null]
8	tt3392194	[null]	[null]
9	tt3392202	[null]	[null]
10	tt3392206	[null]	[null]

Person:

Data Output Explain Messages Notifications

	person_id [PK] character varying	primaryName character varying	photos character varying[]	birthYear bigint	popular_works character varying[]
1	nm0000001	Fred Astaire	[null]	1899	{tt0031983,tt0072308,tt0053137,tt0050419}
2	nm0000002	Lauren Bacall	[null]	1924	{tt0117057,tt0037382,tt0071877,tt0038355}
3	nm0000003	Brigitte Bardot	[null]	1934	{tt0056404,tt0049189,tt0057345,tt0054452}
4	nm0000004	John Belushi	[null]	1949	{tt0080455,tt0078723,tt0072562,tt0077975}
5	nm0000005	Ingmar Bergman	[null]	1918	{tt0050986,tt0050976,tt0060827,tt0069467}
6	nm0000006	Ingrid Bergman	[null]	1915	{tt0034583,tt0038787,tt0038109,tt0077711}
7	nm0000007	Humphrey Bogart	[null]	1899	{tt0033870,tt0042593,tt0034583,tt0043265}
8	nm0000008	Marlon Brando	[null]	1924	{tt0070849,tt0078788,tt0068646,tt0047296}
9	nm0000009	Richard Burton	[null]	1925	{tt0059749,tt0087803,tt0057877,tt0061184}
10	nm0000010	James Cagney	[null]	1899	{tt0029870,tt0031867,tt0035575,tt0042041}

TV Series:

Data Output Explain Messages Notificatic

	series_id [PK] character varying	is_running boolean
1	tt0029270	true
2	tt0030298	true
3	tt0032557	true
4	tt0035599	false
5	tt0035803	false
6	tt0038276	false
7	tt0038309	true
8	tt0038738	true
9	tt0039120	false
10	tt0039121	false

Person_Episodes:

Data Output | Explain | Messages | Notifications

	Person_person_id character varying	Episodes_episode_id character varying	role character varying	character_name character varying[]
1	nm0756803	tt1385036	actor	{Gabriel}
2	nm0940890	tt1385036	actress	{Therese}
3	nm0249103	tt1385036	actor	{Victor}
4	nm0404014	tt1385036	director	[null]
5	nm0196745	tt1385036	writer	[null]
6	nm0384643	tt1385036	writer	[null]
7	nm0253673	tt1385036	producer	[null]
8	nm0008066	tt1385036	production_designer	[null]
9	nm2564961	tt1385059	editor	[null]
10	nm3348851	tt1385059	actress	{"Mary Magdalene"}

Person_Generic_Media:

Data Output | Explain | Messages | Notifications

	Person_person_id character varying	Generic_Media_IMDB_id character varying	role character varying	character_name character varying[]
1	nm0409390	tt0000690	actor	{"Irving Robertson"}
2	nm0813603	tt0000690	actor	{"Frank Wilson"}
3	nm0311375	tt0000690	actor	{"Henderson - the Mana...
4	nm0331049	tt0000690	actor	[null]
5	nm0424530	tt0000690	actor	[null]
6	nm0125509	tt0000791	actor	[null]
7	nm0697944	tt0001065	actor	{"Tim Noonan"}
8	nm0424530	tt0001065	actor	{Policeman}
9	nm0642722	tt0001065	actor	{"Factory Superintenden...
10	nm0409390	tt0001065	actor	{Husband}

Generic_Media_Location:

Data Output | Explain | Messages | Notifications

	Generic_Media_IMDB_id character varying	Location_country character varying	release_title character varying	language character varying
1	tt0000072	US	Officers of French Arm...	[null]
2	tt0000072	FR	Départ des officiers	[null]
3	tt0000159	FR	Le cabinet de Méphist...	[null]
4	tt0000159	US	The Devil's Laboratory	[null]
5	tt0000159	XWW	The Cabinet of Mephis...	en
6	tt0000159	US	The Laboratory of Mep...	[null]
7	tt0000240	US	Delivering Newspapers	[null]
8	tt0000240	US	Distributing a War Extr...	[null]
9	tt0000240	US	World News Wagon	[null]
10	tt0000240	US	Distributing a War Extra	[null]

Episode_Location:

Data Output | Explain | Messages | Notifications

	Episodes_episode_id character varying	Location_country character varying	release_title character varying	language character varying
1	tt0041951	US	The Tenderfeet	[null]
2	tt0070551	US	Slow Boy	[null]
3	tt0075671	US	Spider-Man	[null]
4	tt0076190	PL	Niesamowity Hulk	[null]
5	tt0076190	ES	La masa, un hombre in...	[null]
6	tt0076190	GR	O teratanthropos	[null]
7	tt0076190	IT	L'incredibile Hulk	[null]
8	tt0076190	US	The Incredible Hulk	[null]
9	tt0076190	FI	Hulk - vihreä hurjimus	[null]
10	tt0076190	VE	Hulk, el hombre increíble	[null]