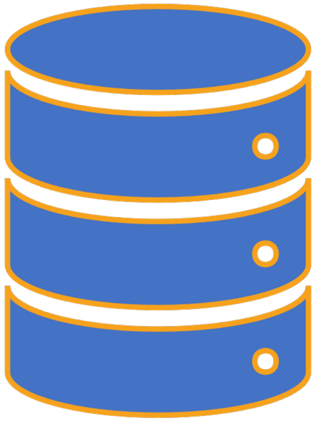


DATABASE DESIGN FOR ASSESSMENT OF PHARMACEUTICAL PRESCRIPTION RISKS

Claryty | SFSU

Nikhit Soares

ROADMAP

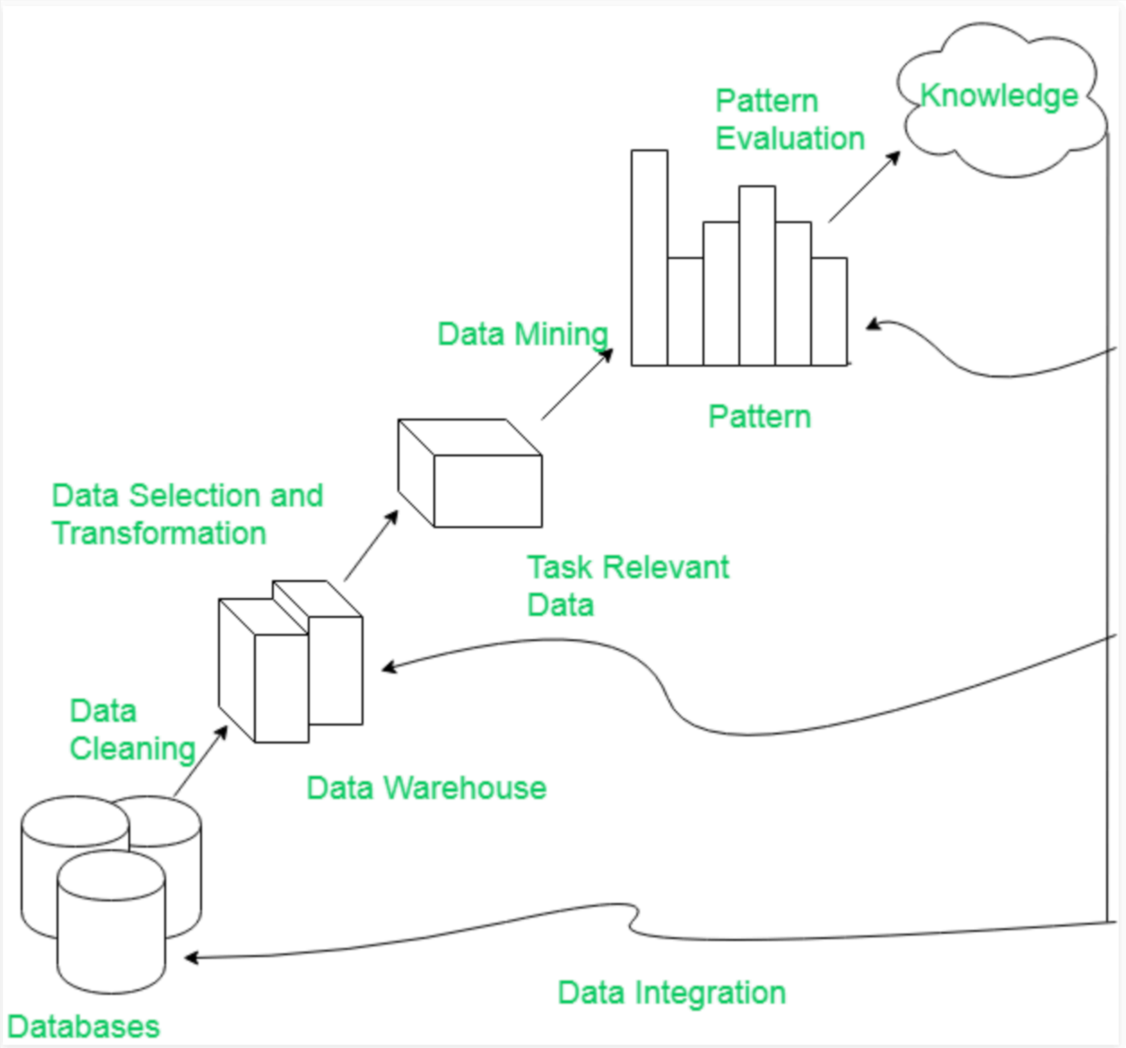
- Background (Claryty, KDD process)
- Introduction to dataset
- Goal and objective
- Challenges
- Methodology (ER diagram, build schema, data pre-processing and loading, data extraction and analysis, query result validation)
- Shift to Python environment(Software change reason, merging columns, data preprocessing, knowledge extraction and analysis)
- Conclusion (Results and Findings, Shortcomings, Future Work)

BACKGROUND

Claryty

- Problem : Limited information to end users pertaining drug quality and risks
- Goal: Develop smart application to showcase end users risk associated with pharmaceutical prescription drugs by providing a risk score.
- Methodology: To collect all risk information associated with pharmaceutical products and showcasing it to end users such that they are easily understood.
- Focus of this project is to build an effective Database management for 'Claryty' and extract useful demographic and drug insights while fulfilling business requirements.
- End-users - Consumers and Healthcare practitioners.

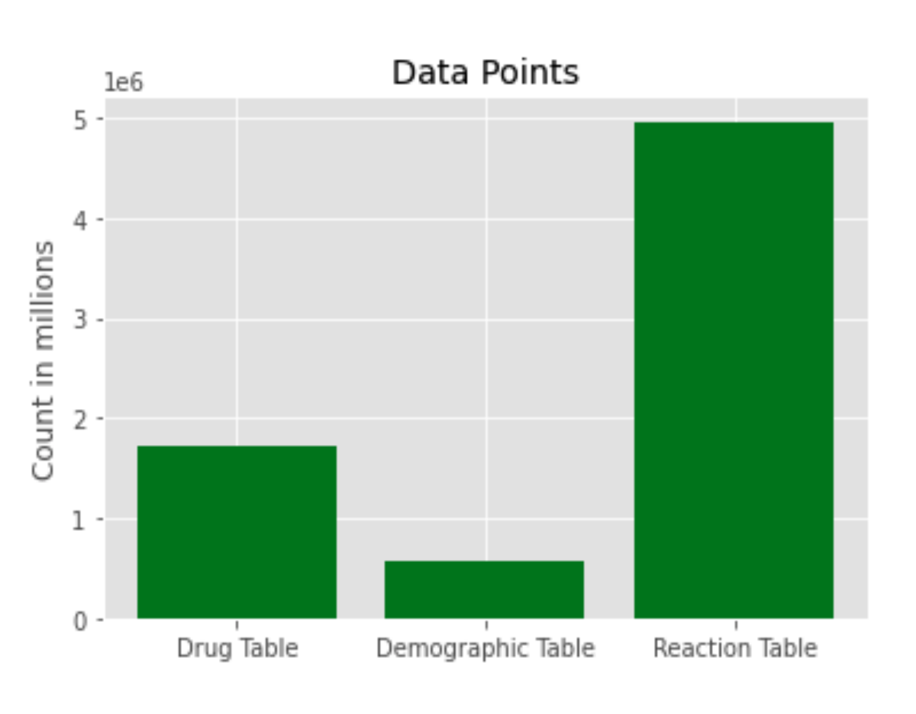
KDD PROCESS



Knowledge
Pattern
Evaluation
Data Mining
Pattern
Data Selection and
Transformation
Task Relevant
Data
Data
Cleaning
Data Warehouse
Databases
Data Integration

FDA ADVERSE EVENT DATA

- https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html
- The quarterly data files are available in ASCII or SGML formats.
- For the scope of the project, research is narrowed to I year(4 quarters) worth data.



ENTITIES AND ATTRIBUTES: DEMOGRAPHIC

- primaryid
- caseid
- caseversion
- i_f_code
- event_dt
- mfr_dt
- init_fda_dt
- mfr_num
- mfr_sndr
- age
- age_cod
- age_grp
- sex
- wt
- wt_cod
- rept_dt
- to_mfr
- occp_cod
- reporter_country
- occr_country

- Contains 20 attributes, capturing patient data like the age, weight and sex.
- Reported country and occurrence country capture the demographics of the patient.
- The table also captures some important dates like, when did the adverse event happen and when did the FDA receive the official report.
- It also contains some vital information about the manufacturer.

ENTITIES AND ATTRIBUTES: DRUG

- primaryid
- caseid
- drug_seq
- role_cod
- drugname
- prod_ai
- val_vbm
- route
- dose_vbm
- cum_dose_chr
- cum_dose_unit
- dechal
- rechal
- exp_dt
- dose_amt
- dose_unit
- dose_form
- dose_freq

- Contains 18 attributes, capturing drug data name and active ingredient.
- Route, dose amount, dose form, and dose freq describes the amount, dose form, and frequency of dose prescribed respectively.
- Interestingly, the table also captures data of the adverse effect reoccurring or not occurring after the drug intake is stopped.
- It also contains the expiry date of the drug.

ENTITIES AND ATTRIBUTES: INDICATION

	primaryid	caseid	indi_drug_seq	indi_pt
0	38941682	3894168	1	Parkinsonism
1	38941682	3894168	7	Parkinsonism
2	38941682	3894168	11	Parkinsonism
3	38941682	3894168	12	Parkinsonism
4	38941682	3894168	13	Parkinsonism

- Contains 4 attributes, capturing drug sequence number for identifying a drug for a case.
- Indi pt describes the preferred medical term describing the indication to use.

ENTITIES AND ATTRIBUTES: REACTION

primaryid	caseid	pt	drug_rec_act
38941682	3894168	Agitation	NaN
38941682	3894168	Akinesia	NaN
38941682	3894168	Blood creatine phosphokinase increased	NaN
38941682	3894168	Drug ineffective	NaN
38941682	3894168	Dysphagia	NaN

- Contains 4 attributes, capturing the reaction to a patient after using the drug.
- pt describes the preferred medical term describing the adverse event.
- Drug rec act is populated with reaction/event information (PT) if/when the event reappears upon re-administration of the drug.

ENTITIES AND ATTRIBUTES: OUTCOME

primaryid	caseid	outc_cod
37624583	3762458	HO
37646134	3764613	HO
37969093	3796909	DE
37969093	3796909	HO
37969093	3796909	OT

- Contains 3 attributes, capturing the outcome of a patient after using the drug.

- In the above example the outcome code represents the following.
  HO: hospitalization
  DE : death
  OT: Other serious(Important medical event)

ENTITIES AND ATTRIBUTES: REPORT SOURCES

primaryid	caseid	rpsr_cod
154313281	15431328	HP
154507833	15450783	HP
154507833	15450783	SDY
157814021	15781402	CSM
157851291	15785129	CSM

- Contains 3 attributes, capturing the source of the reported case.

- In the above example the rspr code represents the following.
  HP: Health Professional
  SDY : Study
  CSM: Consumer

ENTITIES AND ATTRIBUTES: THERAPY

primaryid	caseid	dsg_drug_seq	start_dt	end_dt	dur	dur_cod
1000808590	10008085	1	20130308.0	NaN	NaN	NaN
1000808590	10008085	2	20140930.0	NaN	NaN	NaN
1000808590	10008085	3	20160121.0	NaN	NaN	NaN
1000808590	10008085	4	20160204.0	NaN	NaN	NaN
1000808590	10008085	5	20160229.0	NaN	NaN	NaN

- Contains 7 attributes, capturing the Therapy dates of a patient.
- Dur shows the duration of the therapy reported.

GOAL AND OBJECTIVE

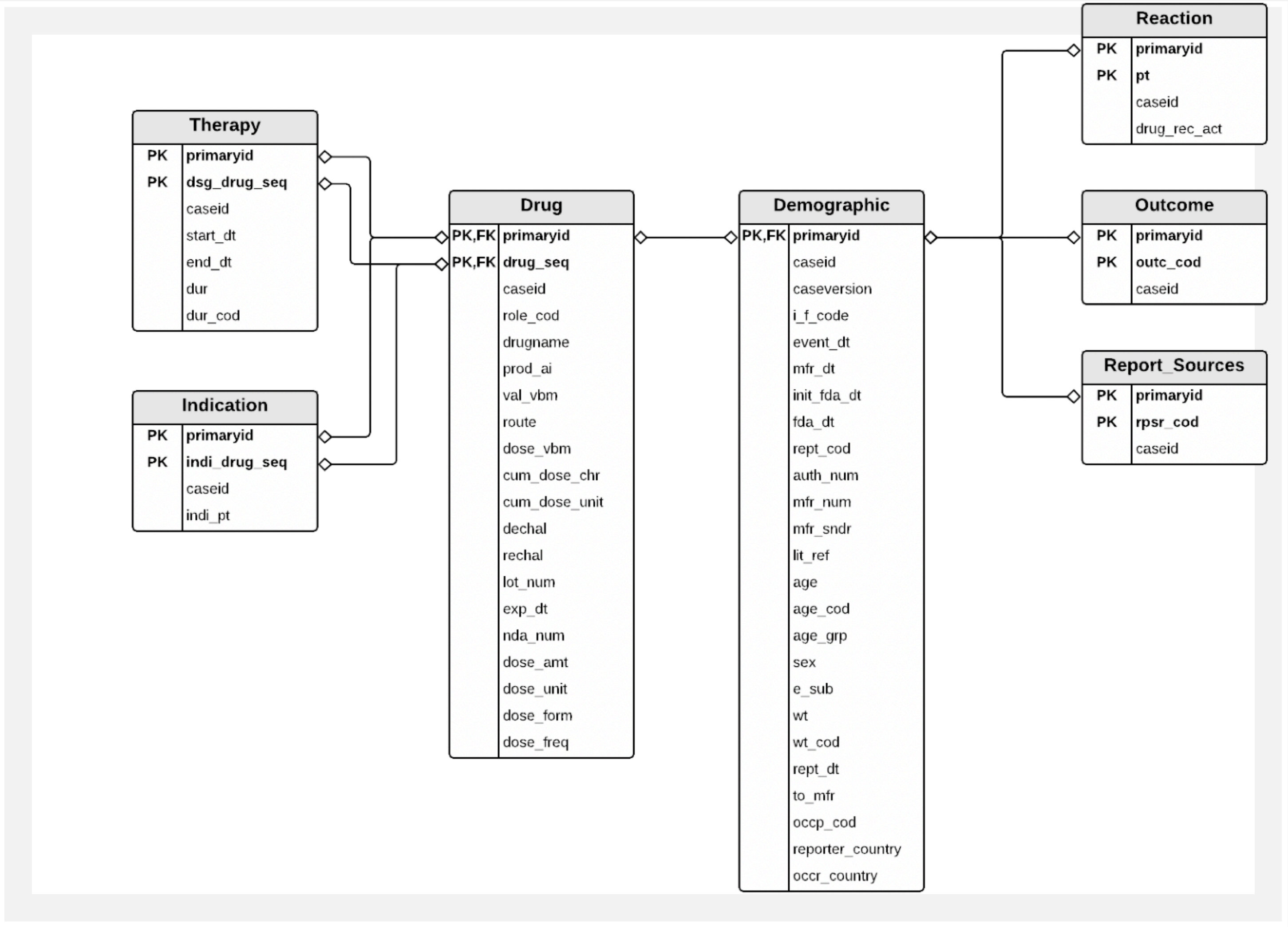
- As mentioned on the FDA Adverse Effect website, the database can be built using SQL.
- One of the main objective of the project is to check whether such a large database can be built on the SQL workbench or not.
- Extracting useful knowledge using robust queries is pivotal.
- Extracting demographics and number of adverse event reports reported for a particular drug with 95% accuracy.
- Return a RxCUI code along with the drug name associated with that particular drug. The RxCUI codes are available online.
- The end output should be in json format to support the input requirements of the application.

CHALLENGES

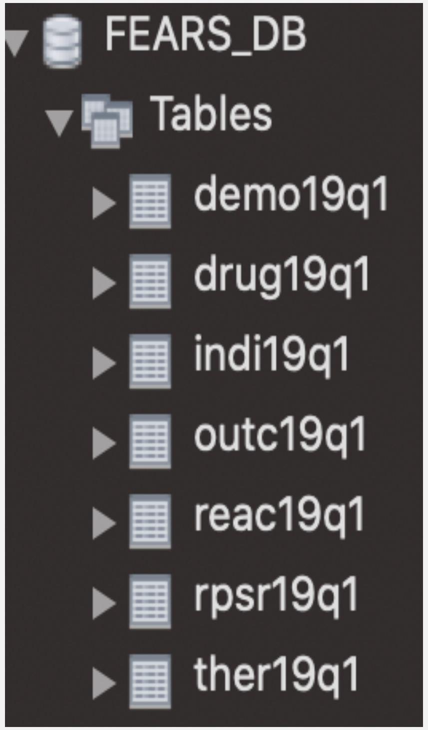
- Large number of null values and duplication. Preprocessing becomes complicated.
- Joining tables.
- Generating robust queries.
- Storing and accessing accurate information.

ER DIAGRAM

	Linked To
Demographic table	Reaction, Outcome , Report Sources (None, one, multiple) Drug(one, multiple)
Drug Table	Indication (None, one, multiple) Therapy(one, many)



PREPROCESSING AND BUILDING DATA SCHEMA



- Change file to acceptable format.
- Sensitive dataset and so can't risk changing or preprocessing most of the fields.
- All the tables are joined with their respective composite primary key and the foreign key as mentioned in the ER Diagram. Data is imported using the Data Import Wizard option on sql workbench.
- A total of 8 columns like literature reference, electronic submission, lot num, etc. are dropped from the demographic and drug table.
- These columns are dropped keeping in mind the business needs and to save query processing time.

KNOWLEDGE EXTRACTION AND ANALYSIS

- By joining drug and 'reaction' tables, it is possible to retrieve information pertaining to a specific drug and their reactions.
- As shown in the figure below, where drugname = 'SELEGLINE' and their symptoms as 'Depression' or 'Dyspnoea'.

```sql
-- Drugname, Active Ingredients, and Symptoms
USE Fears_Test;
SELECT drug19q1.primaryid,drug19q1.drugname AS DRUG, drug19q1.prod_ai AS ACTIVE_INGREDIENT,reac19q1.pt AS MEDICAL_SYMPTOMS
FROM reac19q1
INNER JOIN drug19q1 ON (drug19q1.primaryid=reac19q1.primaryid)
WHERE drug19q1.drugname = 'SELEGILINE'
AND (reac19q1.pt = 'Depression' OR reac19q1.pt = 'Dyspnoea')
;
```

100% 61:26

Result Grid Filter Rows: Search Export:

primaryid	DRUG	ACTIVE_INGREDIENT	MEDICAL_SYMPTOMS
122890185	SELEGILINE	SELEGILINE	Dyspnoea

Result Grid

Form Editor

KNOWLEDGE EXTRACTION AND ANALYSIS

- Number of reports generated for a particular drug is one of the pivotal pieces of information we can assess using the dataset.

- To verify the integrity of the queries a dummy database is made and used. This is one of the most pivotal step while working with a large scaled database.

3894168	1	BROMOCRIPTINE MESYLATE
3894168	2	BROMOCRIPTINE MESYLATE
3894168	7	LEVODOPA+BENSERAZIDE
4022386	1	INFLIXIMAB, RECOMBINANT
3894168	13	SELEGILINE
11259551	3	DICLOFENAC SODIUM.
11468194	1	LIPITOR

```
drugname
BROMOCRIPTINE MESYLATE      2
DICLOFENAC SODIUM.          1
INFLIXIMAB, RECOMBINANT     1
LEVODOPA+BENSERAZIDE        1
LIPITOR                     1
SELEGILINE                  1
Name: drugname, dtype: int64
```

KNOWLEDGE EXTRACTION AND ANALYSIS

- We can also filter results based on Indications and their respective drug.
- This query is tested and then applied on the real database.

```sql
USE Fears_Test;
SELECT drug19q1.primaryid,drug19q1.drugname AS Drug, drug19q1.prod_ai AS Active_Ingredient,indi19q1.indi_pt AS Indications
FROM indi19q1
INNER JOIN drug19q1 ON (drug19q1.primaryid=indi19q1.primaryid)
WHERE drug19q1.prod_ai = 'BOSUTINIB'
AND (indi19q1.indi_pt = 'Hypertension')
;
```

primaryid	Drug	Active_Ingredient	Indications
104825357	BOSUTINIB	BOSUTINIB	Hypertension
104825357	BOSUTINIB	BOSUTINIB	Hypertension
157522172	BOSUTINIB	BOSUTINIB	Hypertension
160243506	BOSUTINIB	BOSUTINIB	Hypertension
160243506	BOSUTINIB	BOSUTINIB	Hypertension
160243506	BOSUTINIB	BOSUTINIB	Hypertension
160243508	BOSUTINIB	BOSUTINIB	Hypertension
160243508	BOSUTINIB	BOSUTINIB	Hypertension
160243508	BOSUTINIB	BOSUTINIB	Hypertension

SHIFT TO PYTHON IDE

- The queries generated on MySql worked perfectly, but it takes 12-15 seconds to return the output.
- The queries are not robust and so the hypothesis of making the database schema on SQL can be negated.
- Shifting to python, the various tables have to be merged using the appropriate joins mentioned in the ER Diagram.
- A python script was already made previously, but now, one one of the main challenges is to check the integrity and accuracy of the script.

# primaryid	caseid	drug_seq	drugname	prod_ai	val_vbm	route	dose_vbm	cum_dose_chr	cum_dose_unit	dechal	rechal	exp_dt	dose_amt	dose_unit	dose_form	dose_freq
1	3894168	1	BROMOCRIPTINE MESYLATE	BROMOCRIPTINE MESYLATE	1	Oral	10 MG, TID	0	NaN	Y	NaN	NaN	10.0	MG	NaN	TID
1	3894168	2	BROMOCRIPTINE MESYLATE	BROMOCRIPTINE MESYLATE	1	Oral	7.5 MG, TID	0	NaN	Y	NaN	NaN	7.5	MG	NaN	TID
1	3894168	7	LEVODOPA+BENSERAZIDE	BENSERAZIDE\LEVODOPA	1	Oral	125 MG, 6QD	0	NaN	NaN	NaN	NaN	125.0	MG	NaN	NaN
4	4022386	1	INFLIXIMAB, RECOMBINANT	INFLIXIMAB	1	Intravenous (not otherwise specified)	WEEKS 0, 2, 6, 10, 18, 26, AND 34 CYCLICAL	0	NaN	U	U	NaN	10.0	MG/KG	SOLUTION FOR INFUSION	NaN
5	3894168	13	SELEGILINE	SELEGILINE	1	Unknown	15 MG, UNK	0	NaN	U	NaN	NaN	15.0	MG	NaN	NaN
6	11259551	3	DICLOFENAC SODIUM.	DICLOFENAC SODIUM	1	NaN	100 MG, ONCE DAILY (QD)	0	NaN	NaN	U	NaN	100.0	MG	NaN	QD
7	11468194	1	LIPITOR	ATORVASTATIN CALCIUM	1	Oral	20 MG, DAILY	0	NaN	U	NaN	NaN	20.0	MG	FILM-COATED TABLET	NaN

MERGING TABLES APPROPRIATELY

- The original script had merged the data frames without considering the ER diagram relationship. Due to this, incomplete data was obtained as the output.
- As seen in the results below, the reactions (pt) of the table is merged with the demographic table. Notice, only unique primary ids are returned.

```
merged_left = pd.merge(left=demoDF, right=reacDF, how='left', left_on='# primaryid', right_on='# primaryid')
merged_left.drop("caseid_y",axis=1,inplace=True)
merged_left
```

# primaryid	caseid_x	caseversion	i_f_code	event_dt	mfr_dt	init_fda_dt	mfr_num	mfr_sndr	age	...	age_grp	sex	wt	wt_cod	rept_dt	to_mfr	occp_cod	reporter_country	occr_country	pt
1	4022386	3	F	0	20191016	20031000	GB-JNJFOC-20031000825	JOHNSON AND JOHNSON	52	...	NaN	M	0	NaN	20191000	NaN	OT	GB	GB	Agitation
2	4076188	6	F	200303	20051109	20040100	PHBS2003JP07309	NOVARTIS	80	...	NaN	M	0	NaN	20190200	NaN	OT	JP	JP	Bile duct obstruction
3	4122546	4	F	0	20040428	20040400	PHBS2004DE04597	NOVARTIS	79	...	NaN	F	0	NaN	20190400	NaN	CN	FR	DE	Cholangitis acute,Decreased appetite,Depression
4	4208085	4	F	20020600	20041020	20040900	PHBS2004JP11897	NOVARTIS	86	...	NaN	F	0	NaN	20190100	NaN	MD	JP	JP	NaN
5	8299276	6	F	20111100	20150312	20111200	DE-ROCHE-1022692	ROCHE	62	...	NaN	F	84	KG	20191200	NaN	MD	DE	DE	NaN
6	8828837	4	F	201203	20190531	20121000	CA-ROCHE-CID000000002173273	ROCHE	54	...	NaN	M	89	KG	20190600	NaN	CN	CA	CA	Duodenal ulcer
7	11806747	4	F	20060500	20191227	20151200	US-PFIZER INC-2015421690	PFIZER	49	...	NaN	M	0	NaN	20191200	NaN	LW	US	US	Dyspnoea

DEMOGRAPHIC ANALYSIS

```
demographic_filtering= merged_left[['# primaryid','age','occr_country','pt']][merged_left['occr_country']=='JP']
demographic_filtering
```

- The results below show filtered records where the occurred country is JP. The age, demographic, and reaction for the particular patients are returned.

# primaryid	age	occr_country	pt
2	80	JP	Bile duct obstruction
4	86	JP	NaN

DATA PREPROCESSING AND KNOWLEDGE EXTRACTION

- Matching drug primary id to demographic id to avoid duplicate records.
- Columns like weight, year and age are converted to consistent unit formats.

```
# Iterating over every row in drugtable and match it with demographic table
merged_df = demoDF1_[demoDF1_.index.isin(drugDF1[drugDF1['drugname'] == drugname]['primaryid'].values)]
print('shape: ',merged_df.shape)
print(merged_df.head())
for index,row in merged_df.iterrows():
    val = row['age']/ageDict[row['age_cod']]
    merged_df.loc[index,'age'] = val
merged_df=merged_df.where(pd.notnull(merged_df), None)
merged_df_= merged_df.to_dict()
```

KNOWLEDGE EXTRACTION

- Get rxCUI code and match with drugname.
- Count the number of adverse reports reported for the requested drug.

```
# Getting the rxCUI code

    rawResponse = requests.get('https://rxnav.nlm.nih.gov/REST/rxcui.json?name=' + drugname)
    res = rawResponse.json()
    if not 'rxnormId' in res['idGroup']:
        merged_df_["id"] = "Not Found"
    else:
        rxCUICode = res['idGroup']['rxnormId'][0]
        merged_df_["id"] = rxCUICode

    merged_df_["drug name"] = drugname
```

```
# Counting length of the returned df to get the number of reports
    merged_df_["No of Reports"] = len(merged_df_['caseid'])
# Converting to json
    json_object = json.dumps(merged_df_, indent = 4)
    with open("sample2.json", "w") as outfile:
        json.dump(merged_df_, outfile)
    print(json_object)
```

KNOWLEDGE EXTRACTION

- According to our dataset, 5 adverse events recorded from drug name Salfasalazine.
- The output shows the extracted information in json format.

primaryid	caseid	drug_seq	drugname	prod_ai
111223355	11122335	7	SALAZOPIRINA	SULFASALAZINE
144785407	14478540	17	SALAZOPIRINA	SULFASALAZINE
159079631	15907963	7	SALAZOPIRINA	SULFASALAZINE
159350762	15935076	3	SALAZOPIRINA	SULFASALAZINE
161660871	16166087	2	SALAZOPIRINA	SULFASALAZINE

```
"id": "Not Found",
"drug name": "SALAZOPIRINA",
"No of Reports": 5
```

```
rxCUI:             "142426"
drug name:         "BROMOCRIPTINE MESYLATE"
No of Reports:     6
```

RESULTS


```
"occp_cod": {
    "111223355": "CN",
    "144785407": "CN",
    "159079631": "CN",
    "159350762": "MD",
    "161660871": "CN"
},
"reporter_country": {
    "111223355": "PT",
    "144785407": "CA",
    "159079631": "PT",
    "159350762": "ES",
    "161660871": "PT"
},
"occr_country": {
    "111223355": "PT",
    "144785407": "CA",
    "159079631": "PT",
    "159350762": "ES",
    "161660871": "PT"
```

```
"event_dt": {
    "111223355": 201504.0,
    "144785407": 20161100.0,
    "159079631": 201504.0,
    "159350762": 20181200.0,
    "161660871": 20181000.0
},
"mfr_dt": {
    "111223355": 20190114.0,
    "144785407": 20190327.0,
    "159079631": 20190131.0,
    "159350762": 20190207.0,
    "161660871": 20190304.0
```

```
"caseid": {
    "111223355": 11122335,
    "144785407": 14478540,
    "159079631": 15907963,
    "159350762": 15935076,
    "161660871": 16166087
},
"caseversion": {
    "111223355": 5,
    "144785407": 7,
    "159079631": 1,
    "159350762": 2,
    "161660871": 1
```

```
"init_fda_dt": {
    "111223355": 20150500.0,
    "144785407": 20180200.0,
    "159079631": 20190200.0,
    "159350762": 20190200.0,
    "161660871": 20190400.0
},
"mfr_num": {
    "111223355": "PT-ABBVIE-15K-130-1390794-00",
    "144785407": "CA-TAKEDA-2016TUS022277",
    "159079631": "PT-PFIZER INC-2019048131",
    "159350762": "ES-ROCHE-2261941",
    "161660871": "PT-ABBVIE-19K-130-2694403-00"
},
```

```
"mfr_sndr": {
    "111223355": "ABBVIE",
    "144785407": "TAKEDA",
    "159079631": "PFIZER",
    "159350762": "ROCHE",
    "161660871": "ABBVIE"
},
"age": {
    "111223355": 68.0,
    "144785407": 61.0,
    "159079631": 68.0,
    "159350762": 56.0,
    "161660871": 62.0
},
"age_cod": {
    "111223355": "YR",
    "144785407": "YR",
    "159079631": "YR",
    "159350762": "YR",
    "161660871": "YR"
```

CONCLUSION

- The database is huge, and so MySql queries take long time to run on local hard-disk.
- The data provided by the FDA is noisy. So, more preprocessing and refining will be needed to create data extraction and visualization.
- The original python script joins tables in a way that the results are not 95% accurate as needed.
- Merging the tables according to the ER diagram can help attain the required accuracy.
- Python queries are robust comparatively.

SHORT COMINGS AND FUTURE WORK

- The dataset is certainly huge, and so storage capacity will be a pivotal concern to ensure robust query execution.
- The dataset is rich with data, which means ton of knowledge is hidden behind.
- For starters, the data contains adverse effects reported by 199 different countries, consisting of approximately 12,052,68 unique reports for year 2019.

REFERENCES

- Y. Ji, F. Shen and J. Tran, "A Multi-relational Association Mining Algorithm for Screening Suspected Adverse Drug Reactions," 2014 11th International Conference on Information Technology: New Generations, Las Vegas, NV, 2014, pp. 407-412, doi: 10.1109/ITNG.2014.96.
- Daniel Foley. "Let's Build a Streaming Data Pipeline", towards data science, May 7, 2019

https://towardsdatascience.com/lets-build-a-streaming-data-pipeline-e873d671fc57

- Sujay B., et al., " Adverse Drug Reaction Detection System On the basis of Clinical Data", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 4, 2016.

THANK YOU