



NYC XPLORE

TEAM- SEMICOLON

Aditi Godbole





MOTIVATION



- New York is a large city and with that, it brings a plethora of options to choose from within restaurants, hotels, Airbnb etc.
- This portal will provide the user with an option to search for these options within a single portal thus making exploring options a little less daunting and the New York experience a little more carefree.



PREVIOUS IMPLEMENTATION



The previous implementation included datasets like- Airbnb, Hotels, ATM's, Art galleries, Theaters etc.

Persisting Problems in previous implementation

- Problems with importing all the tuples of AirBnb Dataset.
- All the datasets dealt with different concepts → Problems for Duplicate detection.
- Problems with the formation of Global Schema
- Schema Matching issues



TABLE OF CONTENTS

01

PROJECT SUMMARY

A brief summary of the project

03

TECHNOLOGIES USED

An overview of tools and technologies used

05

DEMO & VISUALIZATION

Demo of the project & Visualization of sample queries

02

ARCHITECTURE

Architecture type & Heterogeneities present

04

SUPPORTED QUERIES

Sample queries that will be supported by the application

06

IMPLEMENTATION ISSUES

Problems Encountered



01

PROJECT SUMMARY

A brief overview of the project

NEW DATASETS

1. NYC Hotels Dataset (File format- .csv, Source- Kaggle)
2. Time Square Hotels Dataset (File format- .csv, Source- Kaggle)
3. NYC AirBnb Dataset (File format- .csv, Source- Kaggle)
4. NYC Eateries Dataset (File format- .json, Source- Data World)
5. NYC Restaurants Dataset (File format- .json, Source- Data/City of New York)
6. NYC Michelin Star Restaurants (File format-.csv, Source- Kaggle)

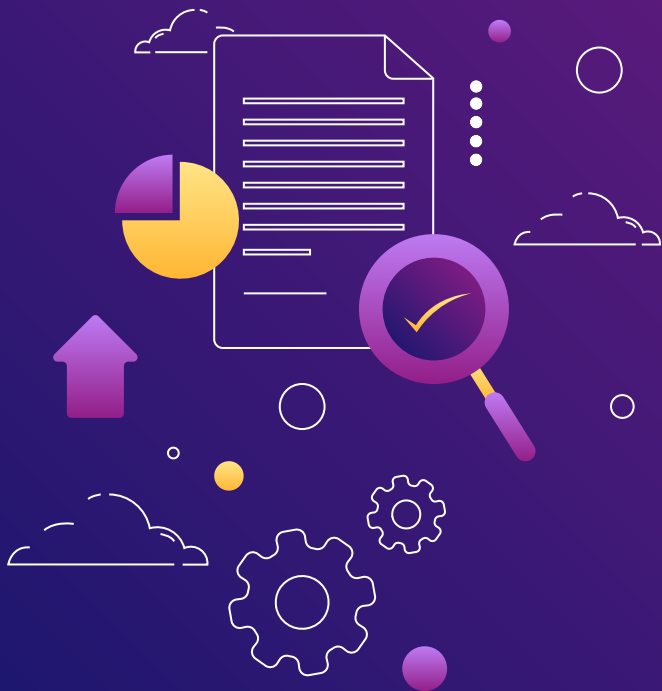


DATASETS : A SNAPSHOT



- 1) NYC Hotels (Rows- 1614, Columns- 8)
Attributes(ean_hotel_id,name,address,city,state_province,postal_code,latitude,longitude)
- 2) NYC AirBnB(Rows- 48895, Columns-8)
Attributes(id,name,host_id,host_name,neighbourhood_group,neighbourhood,latitude,longitude)
- 3) NYC Times Square Hotels (Rows- 41, Columns-10)
Attributes(name,address,phone,website,location,borough,postal_code,latitude,longitude, id)
- 4) NYC Eateries(Rows-237, Columns-8)
Attributes(name,location,description,permit_number,phone,website,type_name,id)
- 5) NYC Restaurants (Rows-21691, Columns- 9)
Attributes(CAMIS, DBA,BORO,building, street, zipcode, phone, cuisine description,id)
- 6) NYC MichelinStar Restaurants(Rows-72, Columns-8)
Attributes(id,name,address,city,state,description,full_address,postal_code)





STANDARDIZATION

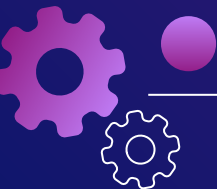
-
- In NYC_Eateries dataset, city column was added and its value was set to New York
- In NYC_TimesSquare dataset, city column was added and its value was set to Manhattan
- In NYC_Eateries, NYC_MichelinStarRestaurants, NYC_Restaurants, A starred column was added and a corresponding TRUE OR FALSE value was set.



GLOBAL SCHEMA- HOTELS

- Target Schema Model- Relational
- Total rows- 50050, Columns- 5
- Attributes(id, Name,Address, Location, Postal_code Latitude, longitude)

	id	Name	Address	Location	Postal_code	Latitude	Longitude
▶	1	The Manhattan at Times Square Hotel	790 7th Ave	New York	10019	40.7621	-73.9826
	2	The Plaza Hotel	Fifth Avenue at Central Park South	New York	10019	40.7642	-73.9739
	3	Residence Inn by Marriott White Plains Westchester County	5 Barker Avenue	White Plains	10601	41.0346	-73.7701
	4	Days Inn Weedsport	9050 Rte 34	Weedsport	13166	43.0561	-76.5587
	5	Fairfield Inn & Suites by Marriott Olean	3270 Route 417	Olean	14760	42.0807	-78.479
	6	Radisson Hotel Hauppauge-Long Island	110 Vanderbilt Motor Pkwy	Hauppauge	11788	40.806	-73.2654
	7	Sheraton New York Times Square Hotel	811 7th Ave	New York	10019	40.7626	-73.9821
	8	Sheraton Syracuse University Hotel & Conference Center	801 University Ave	Syracuse	13210	43.0411	-76.1345
	9	ONE UN New York	One United Nations Plaza	New York	10017	40.7505	-73.9698
	10	Grand Hyatt New York	109 East 42nd Street	New York	10017	40.7517	-73.9766





GLOBAL SCHEMA- RESTAURANTS

- Target Schema Model- Relational
- Total rows- 21990, Columns- 6
- Attributes(id, Name, Address, Location, Postal_Code, Starred)

	id	Name	Address	Location	Postal_code	Starred
	319	Aska	47 S 5th St	Brooklyn	11249	TRUE
	320	L'Appart	225 Liberty St	New York	10281	TRUE
	321	Meadowsweet	149 Broadway	Brooklyn	11211	TRUE
	322	Peter Luger Steak House	178 Broadway	Brooklyn	11211	TRUE
	323	Faro	436 Jefferson St	Brooklyn	11237	TRUE
	324	Blanca	261 Moore St	Brooklyn	11206	TRUE
	325	The River CafÃ©	1 Water St	Brooklyn	11201	TRUE
	326	La Vara	268 Clinton St	Brooklyn	11201	TRUE
	327	The Finch	212 Greene Ave	Brooklyn	11238	TRUE
	383	NOTARO RESTAURANT	635SECOND AVENUE	MANHATTAN	10016	FALSE
	384	AKIMOTO SUSHI	187CHURCH STREET	MANHATTAN	10007	FALSE
	386	"W" CAFE	3905TH AVE	MANHATTAN	10018	FALSE
	387	TOMOE SUSHI	172THOMPSON STREET	MANHATTAN	10012	FALSE
	388	SALUMERIA BEILLESE/ BIRICCHINO REST	3788 AVENUE	MANHATTAN	10001	FALSE
	389	FAMOUS FAMIGLIA PIZZERIA	1398MADISON AVENUE	MANHATTAN	10029	FALSE
	390	O'NIEALS	174GRAND STREET	MANHATTAN	10013	FALSE





DUPLICATE DETECTION- RESTAURANTS & HOTELS

Problems Encountered-

- Multiple rows had same values for all attributes in Restaurants Global Schema and hence had to be deleted thus reducing the number of tuples to 6967.
- Since, a lot of restaurants were chain restaurants, duplicates could not be detected based on only Name attribute.
- In hotels global schema, the Airbnb dataset consisted of NaN values for Address and Postal_Code Attribute. This cause a problem while parsing the file for calculating tf-idf on Address attribute.

Methods used for Duplicate Detection-

- Tf-IDF
- Cosine Similarity



DUPLICATE DETECTION- RESTAURANTS- EG



```
df.loc[row[191]]
```

```
id 185
Name east river park food cart
Address east river park, near tennis courts
Location New York
Postal_code NaN
Starred False
Frequency 1
Address1 east river park near tennis courts
tokenized_Names [east, river, park, food, cart]
tokenized_Address [east, river, park, near, tennis, courts]
Name: 183, dtype: object
```

```
[63] df.loc[row[192]]
```

```
id 186
Name east river park food cart
Address east river park
Location New York
Postal_code NaN
Starred False
Frequency 1
Address1 east river park
tokenized_Names [east, river, park, food, cart]
tokenized_Address [east, river, park]
Name: 184, dtype: object
```



DUPLICATE DETECTION- HOTELS- EG



Duplicate Rows based on 'Name' column are:

	Name	Address	...	Longitude	id
215	Americas Best Value Inn	755 Smithtown Bypass	...	-73.1625	216
314	Econo Lodge Inn And Suites	528 Route 3	...	-73.5002	315
319	Econo Lodge	2303 N Triphammer Rd	...	-76.4852	320
356	Econo Lodge	1449 State Route 9	...	-73.6997	357
424	Quality Inn	551 S Transit St	...	-78.6969	425
439	Rodeway Inn & Suites	1951 Niagara Falls Blvd	...	-78.8222	440
467	Quality Inn	100 Spring Valley Marketplace	...	-74.0256	468
570	Americas Best Value Inn	19 Booth Drive	...	-73.4982	571
603	Americas Best Value Inn	6037 Route 96	...	-77.3539	604
660	Americas Best Value Inn	196 South Hamilton St.	...	-77.1021	661
739	Days Inn	1120 Niagara Falls Blvd	...	-78.8229	740
762	Quality Inn	2788 Hamburg Street	...	-73.9366	763
803	Americas Best Value Inn	473 Hamilton St	...	-77.0034	804
875	Quality Inn & Suites	114 State Route 28	...	-74.0316	876
963	Quality Inn	4142 Albany Post Rd	...	-73.9307	964

[15 rows x 7 columns]



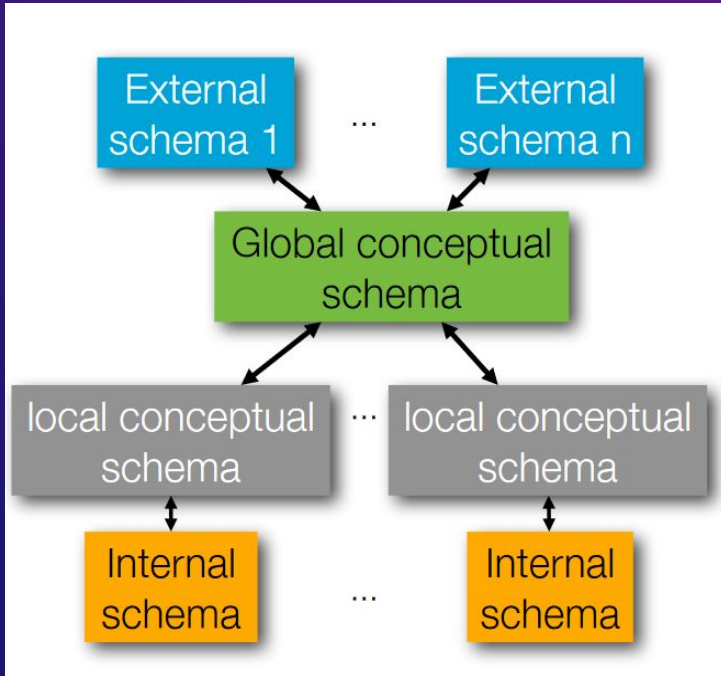
02

ARCHITECTURE

Architecture type & Heterogeneities present



ARCHITECTURE: 4 LEVEL ARCHITECTURE



- Internal Schema-
Refers to the storage medium of Databases.
- Local conceptual Schema-
The local conceptual schema includes all the source hotel and restaurants dataset.
- Global Conceptual Schema-
The global schema comprises of integrated schemas like in this case- global schema hotels & global schema restaurants.
- External Schema 'n'-
It will provide a view of application-relevant data only



HETEROGENEITIES OBSERVED



The following types of heterogeneities were observed:-

1) Semantic Heterogeneity (Name Conflicts- Synonyms)-

Eg- In the eateries dataset, name attribute was represented as name while in restaurants attribute, name was present as DBA.

2) Schematic Heterogeneity (Structural Conflicts- Missing Attributes)-

Eg- In the airbnb dataset, the values of Name attribute were missing for few tuples.

3) Semantic Heterogeneity (Identity)-

Eg- Duplicates present across the integrated tables.

4) Semantic Heterogeneity (Value)-

Eg- Duplicates had different attribute values for semantically equivalent attributes.



03

TECHNOLOGIES USED

An overview of tools and technologies used





DIFFERENT TECHNOLOGIES USED



For Backend:

- Python
- SQL

Libraries:

- pandas
- numpy
- NLTK
- sklearn

For Visualization:
Tableau



04

SUPPORTED QUERIES

Sample queries that will be supported by the application





SUPPORTED QUERIES- QUERY 1

INFORMATION INTEGRATION

- What are the hotels in postal code xyz or abc?

Query- SELECT Name, Postal_code FROM iiproj.globalschema_hotels
WHERE Postal_code= 10017 OR Postal_code = 10591;

	Name	Postal_code
▶	ONE UN New York	10017
	Grand Hyatt New York	10017
	Westchester Marriott	10591
	InterContinental New York Barday	10017
	The Lexington New York City, Autograph Collection	10017
	DoubleTree by Hilton Tarrytown	10591
	Residence Inn by Marriott New York Manhattan/Midtown East	10017
	New York Marriott East Side	10017
	Hilton Manhattan East	10017
	The Roosevelt Hotel, New York City	10017
	Courtyard by Marriott Tarrytown Greenburgh	10591
	Fitzpatrick Grand Central	10017
	The Library Hotel by Library Hotel Collection	10017
	Dylan Hotel NYC	10017
	The Roger Smith Hotel	10017
	Beekman Tower by Bridgestreet	10017
	SpringHill Suites by Marriott Tarrytown/Greenburgh	10591

34 row(s) returned



NYC XPLORE- A portal to explore restaurants and hotels





SUPPORTED QUERIES- QUERY 2

INFORMATION INTEGRATION

- What hotels are located in the same locality(Postal code) as Restaurant xyz?

Query- `SELECT h.Name From iiproj.globalschema_hotels h,iiproj.globalschema_restaurants_copy r
WHERE r.Postal code= h.Postal code AND r.Name ='PANINI GRILL';`

Result Grid		 Filter Rows: <input type="text"/>	Export: 	Wrap
	Name			
▶	Lotte New York Palace			
	Omni Berkshire Place			
	Waldorf Astoria New York			
	The Lombardy			
	San Carlos Hotel			
	W New York			
	Four Seasons Hotel New York			
	Pod 51			
	Courtyard by Marriott New York City Manhattan Midtown East			
	Hotel Elysee by Library Hotel Collection			
	Fitzpatrick Manhattan Hotel			
	DoubleTree by Hilton Metropolitan - New York City			
	The Towers of the Waldorf Astoria New York			
	The Sherry Netherland			
	The Kimberly Hotel & Suites			
	The Benjamin			
	Fifty NYC-an Affinia hotel			
	Renaissance New York Hotel 57			
	The St. Regis New York			

24 row(s) returned



NYC XPLORE- A portal to explore restaurants and hotels



SUPPORTED QUERIES- QUERY 3

INFORMATION INTEGRATION

- List all the Marriott Hotels in NYC

Query- SELECT Name, Location FROM iiproj.globalschema_hotels
WHERE Name like "%Marriott%" AND Location= 'New York';

Name	Location
JW Marriott Essex House New York	New York
Residence Inn by Marriott New York Manhattan/Midtown East	New York
New York Marriott Marquis	New York
New York Marriott Downtown	New York
Courtyard by Marriott New York City Manhattan Fifth Avenue	New York
New York Marriott East Side	New York
Courtyard by Marriott New York City Manhattan Midtown East	New York
Residence Inn by Marriott New York Manhattan/Times Square	New York
Courtyard by Marriott New York Manhattan/Upper East Side	New York
Fairfield Inn by Marriott New York Manhattan/Times Square	New York
Marriott Vacation Club Pulse, New York City	New York
Fairfield Inn by Marriott New York Manhattan/Fifth Avenue	New York
Courtyard by Marriott New York Manhattan/Times Square	New York
Courtyard by Marriott New York Manhattan/SoHo	New York
Fairfield Inn & Suites by Marriott New York Manhattan/Chelsea	New York
Courtyard by Marriott New York Manhattan/Herald Square	New York
SpringHill Suites by Marriott New York Midtown Manhattan	New York
Courtyard by Marriott New York Manhattan / Central Park	New York

20 row(s) returned



NYC XPLORE- A portal to explore restaurants and hotels



SUPPORTED QUERIES- QUERY 4

INFORMATION INTEGRATION

- What Restaurants are located in “Manhattan” ?

Query- SELECT Name from iiproj.globalschema_restaurants_copy
WHERE Location= 'Manhattan'

	Name
▶	NOTARO RESTAURANT
	AKIMOTO SUSHI
	"W" CAFE
	TOMOE SUSHI
	SALUMERIA BEILLESE/ BIRICCHINO REST
	FAMOUS FAMIGLIA PIZZERIA
	O'NIEALS
	DA MIKELE
	RENAISSANCE HOTEL
	PIZZETTERIA BRUNETTI
	DELIMARIE
	AUX EPICES
	YAKITORI TAISHO
	KITCHEN PROVANCE
	NOM WAH TEA/DIM SUM PALOR
	EL BARRIO RESTAURANT
	BALLY TOTAL FITNESS JUICE BAR
	MIDNIGHT EXPRESS

6605 row(s) returned



NYC XPLORE- A portal to explore restaurants and hotels



SUPPORTED QUERIES- QUERY 5

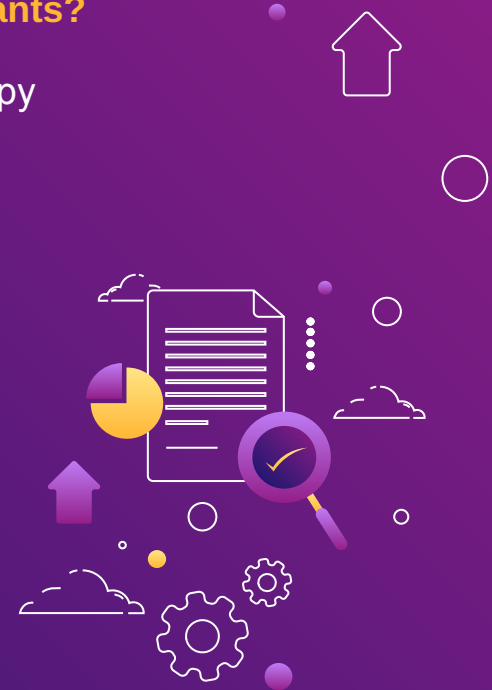
INFORMATION INTEGRATION

- Which restaurants located in New York are Michelin Star Restaurants?

Query- `SELECT Name, Location From iiproj.globalschema_restaurants_copy
WHERE Starred= 'True' AND Location=' New York';`

Result Grid		Filter Rows:	Exp
	Name	Location	
▶	Sushi Inoue	New York	
	Dovetail	New York	
	Cafe Boulud	New York	
	Jean-Georges	New York	
	Masa	New York	
	Per Se	New York	
	Marea	New York	
	Daniel	New York	
	Torishin	New York	
	Le Bernardin	New York	
	The Modern	New York	
	Aquavit	New York	
	Caviar Russe	New York	
	Satsuki (Suzu...	New York	
	Chef's Table ...	New York	
	Aureole	New York	
	Gabriel Kreut...	New York	
	Agern	New York	

62 row(s) returned



NYC XPLORE- A portal to explore restaurants and hotels



SUPPORTED QUERIES- QUERY 6

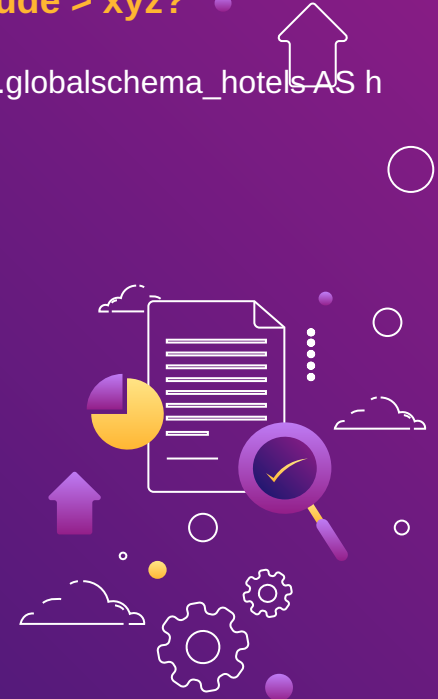
INFORMATION INTEGRATION

- What are the restaurants in the same location as the hotels with latitude > xyz?

Query- SELECT r.Name, Latitude FROM iiproj.globalschema_restaurants_copy AS r, iiproj.globalschema_hotels AS h
WHERE h.Location= r.Location AND h.Latitude > 40.80;

Name	Latitude
Maurice A Fitzgerald Playground Mobile...	40.8018
THE NY PICNIC COMPANY	40.8018
PS 173 Playground Mobile Food Truck	40.8018
NANDITA, INC.	40.8018
Rufus Playground Mobile Food Trucks	40.8018
THE NEW YORK PICNIC COMPANY, INC.	40.8018
Forest Park Mobile Food Truck	40.8018
Battery Park Snack Bar	40.8018
ALEX LINDOWER PARK TRUCKS	40.8018
Cross Island Mobile Food Truck	40.8018
Williamsbridge Oval Park	40.8018
Peter Minuit Snack Bar	40.8018
Alexander Hamilton Plygrd Mobile Food ...	40.8018
Prospect Park Mount Food Cart	40.8018
First Park Outdoor Cafe	40.8018
CAFÉ PRODUCTS CORP.	40.8018
Fort Tryon Park Mobile Food Truck	40.8018

2410 row(s) returned



NYC XPLORE- A portal to explore restaurants and hotels



05

DEMO & VISUALIZATION

Demo of the project & Visualization of sample queries





06

IMPLEMENTATION ISSUES



Problems Encountered



IMPLEMENTATION ISSUES

- A presence of NaN fields.
- Presence of Heterogeneities.
- Initially, it was possible to import only few records from Airbnb due to encoding issues. This issue was handled by populating the table
- Problems while integrating into Global Schema due to missing columns in local schema.
- Presence of tuples having same attribute values across all fields i.e Data Repetition.
- Problems while converting Address attribute to tf-idf values due to missing values.
- Visualization of all queries was not possible since query 2 & query 4 were returning values only for one column



REFERENCES

- https://hpi.de/fileadmin/user_upload/hpi/navigation/10_forschung/30_publicationen/15_dissertationen/Diss_Bleiholder.pdf
- https://help.tableau.com/current/pro/desktop/en-us/examples_mysql.htm
- https://ilias3.uni-stuttgart.de/goto_Uni_Stuttgart_file_2676576_download.html
- <https://towardsdatascience.com/finding-word-similarity-using-tf-idf-in-a-term-context-matrix-from-scratch-in-python-e423533a407>
- <http://www.ultravioletanalytics.com/blog/tf-idf-basics-with-pandas-scikit-learn>
- <https://itectec.com/database/thesql-delete-duplicates-leaving-one-of-two-identical-records/>
- <https://stackoverflow.com/questions/53754234/creating-a-tfidfvectorizer-over-a-text-column-of-huge-pandas-dataframe>
- https://colab.research.google.com/github/goodboychan/goodboychan.github.io/blob/main/_notebooks/2020-07-17-04-TF-IDF-and-similarity-scores.ipynb#scrollTo=FuEsB1iG_7KN



THANK YOU!

