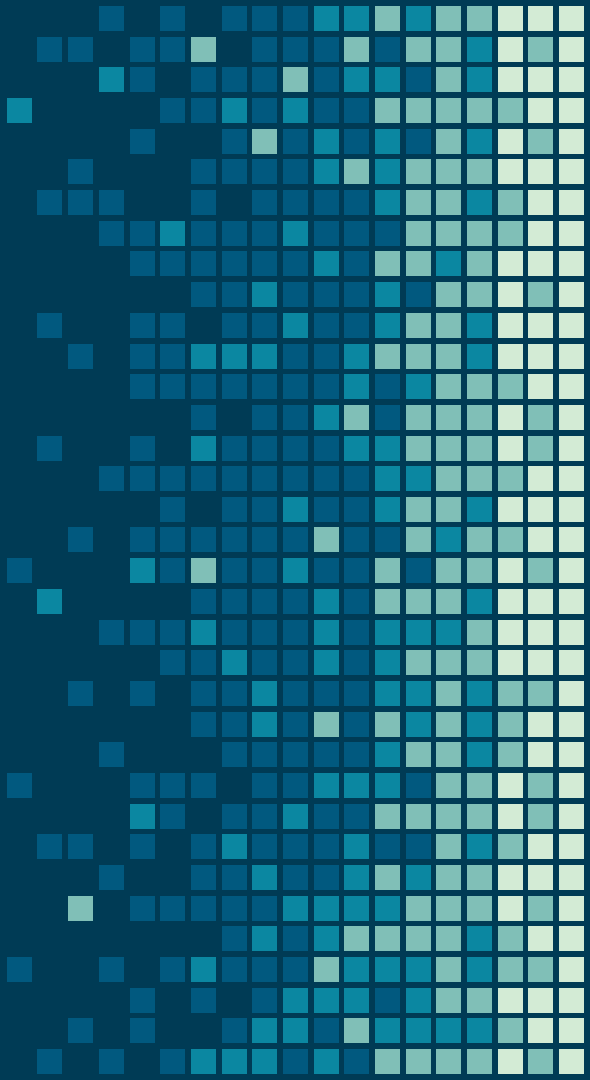
NYC RESTUARANT INSPECTIONS

AIRAGHI DAVIDE

Master Business Intelligence and Big Data Analytics
Università degli studi Milano Bicocca
a.a 2020 - 2021



Agenda

1) **CONTESTO DI RIFERIMENTO**
2) **OBBIETTIVO**
3) **ARCHITETTURA**
4) **DONE / TO DO**
5) **HINT PER SVILUPPI FUTURI**

Contesto

ISPEZIONI SANITARIE AI RISTORANTI A NEW YORK (Restaurant Grading program)

Il dato utilizzato durante il lavoro fa riferimento alle ispezioni sanitarie ai ristoranti da parte del dipartimento di igiene della città di NewYork.

Le ispezioni vengono effettuate circa una volta all' anno e una volta terminato il processo, il ristorante riceve una lettera (A,B,C…) che sarà obbligato ad esporre in bella vista all' ingresso.

L' obiettivo é permettere al consumatore di selezionare il ristorante dove é più sicuro mangiare evitando così intossicazioni alimentari o altre patologie derivanti dal consumo di cibo in condizioni sanitarie non appropriate.

Grading:

A: é la valutazione più alta e viene ricevuta solamente se il ristorante rispetta totalmente tutte le dispozioni.

B: il ristorante risulta pulito ma ci sono delle problematiche minori da sistemare

C: il ristorante non risulta a norma e la contaminazione dei cibi é altamente probabile

Benefit derivanti dal programma:

- Riduzione del numero di problemi di salute legati a contaminazione alimentare dei clienti che fanno scelte più consapevoli
- Possibilità di aumento del volume di clienti da parte di ristoranti a norma di legge
- Risparmio da parte dell' amministrazione pubblica per il sostenimento di cure sanitarie legate a contaminazione alimentare

Obiettivo

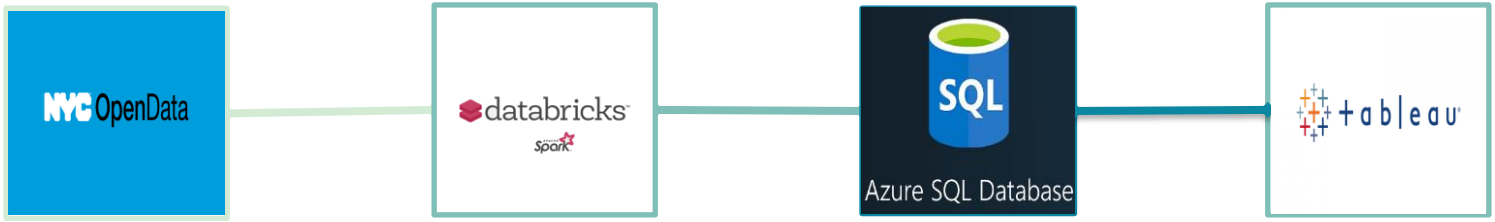
L’obiettivo principale del progetto è predisporre un’ architettura per raccogliere il dato grezzo, permettere di effettuare un’ analisi da differenti prospettive e fornire una visualizzazione del fenomeno per quel che concerne l’ iniziativa di «restaurant grading».

Tasks

I task da eseguire sono:

1. Esplorazione dataset
2. Profilazione e pulizia del dataset
3. Modellazione logica del dato
4. Creazione dimensioni e fatti
5. Creazione del DWH su Azure e caricamento del dato
6. Creazione visualizzazione

Architettura



1° - Dataset

Il dataset utlizzato proviene dal sito ufficiale della città di NewYork inerente la pubblicazione di dati relativi a vari aspetti della vita cittadina.

dataset

2° - ETL

I task di data exploration, data cleaning ed ETL sono eseguiti tramite Jupiter notebook supportato dal motore di analisi Apache Spark in ambiente DataBricks.

databricks_notebook

3° - Datawarehouse

Il dato elaborato viene caricato manualmente su un database SQL Server in ambiente cloud Azure

4° - Data Viz

Il dato viene estratto dal db SQL Server con il fine di creare la visualizzazione su Tableau

NCY_Restaurants_Inspections_viz

Dataset

Il dataset proviene dal sito ufficiale del progetto Open Data di New York e vien aggiornato e alimentato dal DOHMH (Department of Health and Mental Hygiene).

- Il dataset utilizzato è in formato csv ma c' è anche la possibilità di richiamare un' API
- Essendo un dataset della prodotto dalla pubblica amministrazione mi aspetto che non ci siano grossi problemi qualitativi. Oltre a questo sono dati diversi documenti esplicativi a corredo del dataset.
- Al momento dell' estrazione il dataset contiene 400k records ed é strutturato su 26 campi
- Il valore 01/01/1900 nel campo INSPECTION DATE ha un particolare significato ovvero che il ristorante non è ancora stato ispezionato una volta.
- All' interno del dataset sono presenti solamente dati afferenti a ristoranti ancora in attività al momento dell' estrazione
- Il dataset contiene tutte le ispezioni fatte in un certo istante presso un certo ristorante ma bisogna evidenziare che se il ristorante ha riportato più di una violazione il record viene allora duplicato tante volte quante sono le violazioni individuate. Per individuare quindi una singola violazione bisogna quindi accedere per INSPECTION DATE e CAMIS (identificativo del ristorante)
- Qui il link per il dataset

Dataset

Descrizione dei campi considerati

CAMPO	DESCRIZIONE	TIPO
CAMIS	Identificativo univoco del ristorante	string
DBA	Nome del ristorante	string
BORO	Distretto del ristorante (Manhattan,Bronx,etc,)	string
BUILDING	Numero civico	string
STREET	Via	string
ZIPCODE	Codice di avviamento postale	string
PHONE	Telefono	string
CUISINE DESCRIPTION	Tipo di cucina offerta (American,Chinese etc...)	string
INSPECTION DATE	Data ispezione	datetime
ACTION	Azioni intraprese da DOMH (• Violations were cited in the following area(s). • No violations were recorded at the time of this inspection. • Establishment re-opened by DOHMH • Establishment re-closed by DOHMH • Establishment Closed by DOHMH. Violations were cited in the following area(s) and those requiring immediate action were addressed. • "Missing" = not yet inspected)	string
VIOLATION CODE	Codice dell' eventuale violazione	string
VIOLATION DESCRIPTION	Descrizione dell' eventuale violazione	string
CRITICAL FLAG	Indicatore di violazione critica Critical • Not Critical • Not Applicable	
SCORE	Punteggio totale ricevuto al termine dell' ispezione	number
GRADE	Grado associato all' ispezione (N = Not Yet Graded• A = Grade A• B = Grade B• C = Grade C• Z = Grade Pending• P= Grade Pending issued on re-opening following an initial inspection that resulted in a closure)	string
INSPECTION TYPE	Tipo dell' ispezione (riciclo)	string
LATITUDE	Latitudine	number
LONGITUDE	Longitudine	number

ETL

Il mio obbiettivo in questa fase è mettere in qualità il dataset e creare i csv che andranno ad alimentare le dimensioni e la fact table all' interno del datawarehouse su Azure

Svolgo tutti i task di ETL (tramite Python e SQL,API) utilizzando il motore di calcolo Spark in ambiente Databricks che sottometterà tutte le operazioni al cluster temporaneo.

- Data exploration: utilizzo SQL per esplorare un po' il dataset e capire cosa c' è dentro. Inoltre all' interno del notebook c' è la possibilità di creare dei chart istantaneamente sulla base delle query create dandomi così la possibilità di avere un insight immediato su, per esempio, il dominio dei dati che sto osservando.
- Alloco l' intero dataset ad un pandas dataframe ed eseguo quindi la maggior parte delle trasformazioni con le funzionalità messe a disposizione dalla libreria
- I campi VIOLATION CODE e VIOLATION DESCRIPTION riportano le violazioni delle ispezioni. Per alcuni codici (**15F1,15G7,17A3.....)** non è presente una descrizione. Ho chiesto all' owner del dato ma mi ha fornito un pdf (https://www1.nyc.gov/assets/doh/downloads/pdf/rii/blue-book.pdf) che non contiene questi codici che penso non siano ancora stati censiti.
  Ho optato quindi per raggruppare in un' unica categoria ('Other violations') tutte le violazioni per cui non esiste una descrizione.
- Per creare le chiavi surrogate delle dimensioni ho di fatto bisogno di codificare le variabili categoriche in numeri. Utilizzo la funzione Label Encoder della libreria sklearn che assegna un intero in modo randomico senza quindi dare un particolare peso ai valori all' interno della feature. L' unico punto di attenzione è che LE non gestisce i valori null ma il dominio delle variabili su cui lavoro mi permetto di fare delle considerazioni sui valori mancanti e di gestire quindi questa problematica.
- La dimensione TEMPO viene alimentata con una funzione creata ad hoc.

ETL

Voglio dimensionare la mia analisi anche rispetto all' ubicazione dei ristoranti ma rilevo che alcuni attributi fondamentali come zipcode e coordinate geografiche sono mancanti ma la via è sempre presente e conosco per forza di cose la città e lo stato. Installo allora il package googlemaps e utilizzo l' API GeoCoding di Google che dato un indirizzo restituisce un json contenente tutte le informazioni geografiche di un determinato luogo.


```
{   'address_components': [   {   'long_name': '45',
                                  'short_name': '45',
                                  'types': ['street_number']},
                              {   'long_name': 'Rockefeller Plaza',
                                  'short_name': 'Rockefeller Plaza',
                                  'types': ['route']},
                              {   'long_name': 'Manhattan',
                                  'short_name': 'Manhattan',
                                  'types': [   'political',
                                               'sublocality',
                                               'sublocality_level_1']},
                              {   'long_name': 'New York',
                                  'short_name': 'New York',
                                  'types': ['locality', 'political']},
                              {   'long_name': 'New York County',
                                  'short_name': 'New York County',
                                  'types': [   'administrative_area_level_2',
                                               'political']},
                              {   'long_name': 'New York',
                                  'short_name': 'NY',
                                  'types': [   'administrative_area_level_1',
                                               'political']},
                              {   'long_name': 'United States',
                                  'short_name': 'US',
                                  'types': ['country', 'political']},
                              {   'long_name': '10111',
                                  'short_name': '10111',
                                  'types': ['postal_code']}],
    'formatted_address': '45 Rockefeller Plaza, New York, NY 10111, USA',
    'geometry': {   'location': {'lat': 40.7587402, 'lng': -73.9786736},
                    'location_type': 'ROOFTOP',
                    'viewport': {   'northeast': {   'lat': 40.76008918029149,
                                                     'lng': -73.9773246197085},
                                    'southwest': {   'lat': 40.7573912197085,
                                                     'lng': -73.98002258029915}}},
```


E' un metodo di imputazione che però ha dei costi

01–23 ago 2020

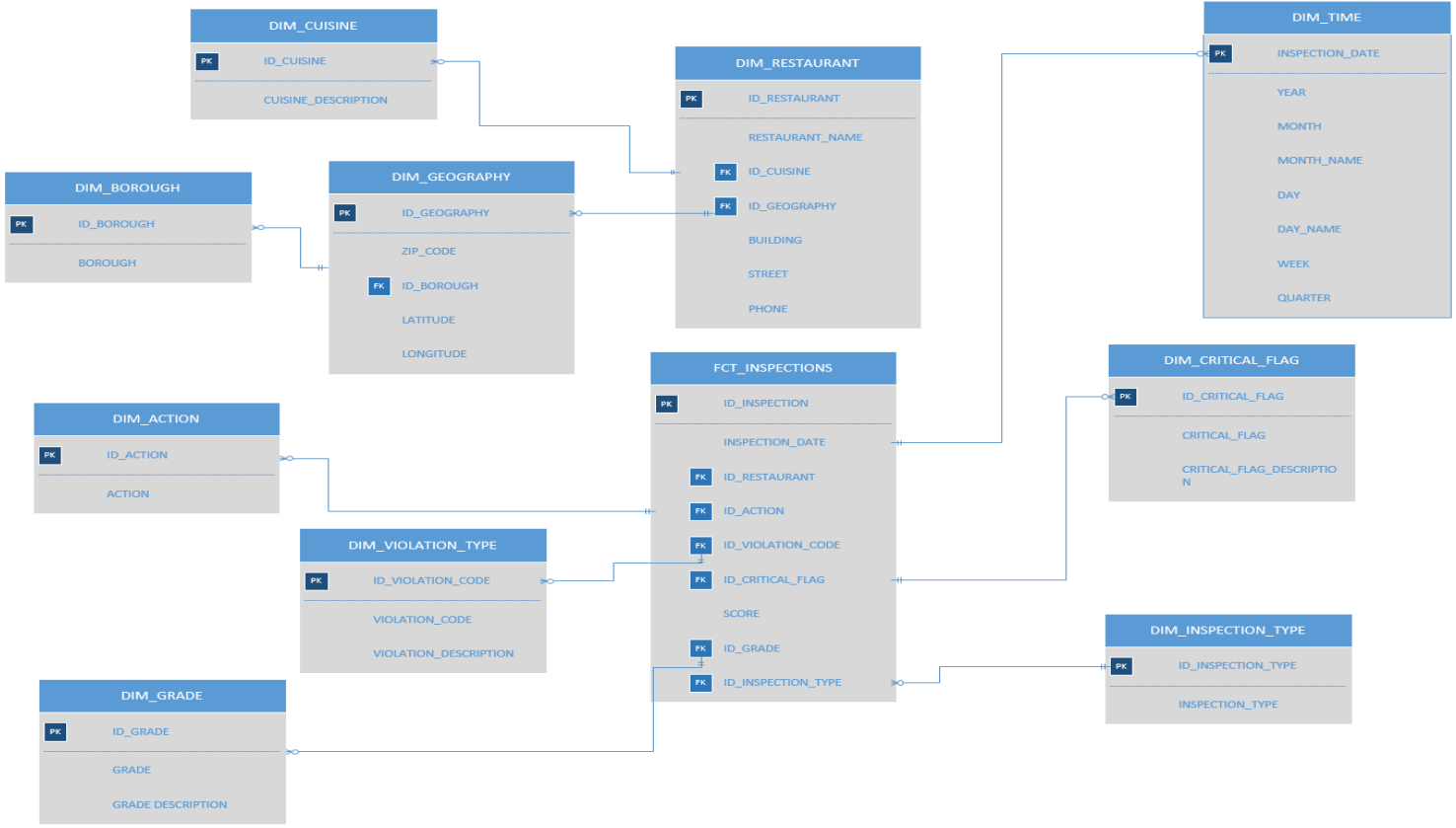
Saldo finale: 0,00 €

Data	Descrizione	Importo (EUR)
01–31 ago 2020	Geocoding API Geocoding: 20041 conteggi [Conversione di valuta: da USD a EUR con un tasso pari a 0.85]	85,21 €

DataWarehouse

- Il dato prodotto dal processo di ETL viene quindi storato su un database SQL Server di Azure (modellato fisicamente come un datawarehouse).
- Avendo la risorsa su Cloud posso disporre della visualizzazione su Tableau con un alto tasso di availability (senza appoggiarmi alla mia macchina locale) e potendo quindi fare un refresh live dei dati.
- Il notebook è di fatto creato per alimentare la tabella dei fatti e le sue dimensioni pertanto se il progetto non fosse one-shot ma fosse in produzione può essere utilizzato per eseguire periodicamente il refresh delle suddette strutture ogni qualvolta che il dataset viene aggiornato.
- Il DWH è alimentato tramite i csv creati nello step precedente e il tutto viene fatto attraverso il wizard di importazione di flat file di Microsoft SQL Server Management Studio.

DataWarehouse



Data Visualization 

Il tool che ho scelto per la visualizzazione é Tableau.

- Stabilisco la connessione con la visualizzazione tramite apposito connettore già built in per Microsoft SQL Server.
- Creo il modello dati importando le strutture presenti sul DWH. Non ho inserito vincoli di integrità referenziale perché sarà il modello del dato di Tableau a stabilire questi vincoli.
- Una volta stabilita la connessione con il dwh posso scegliere di leggere il dato live o creare un' estrazione di tableau che altro non è che uno snapshot del dato ottimizzato per fornire risposte rapide alle query. Ho scelto l' estrazione poiché non ci sono ulteriori caricamenti del dwh quindi il dato al momento rimane statico.
- Qui il link per la visualizzazione: NCY_Restaurants_Inspections_viz

Insights

- La maggior parte dei ristoranti presenti nell' area metropolitana di New York presenta un grading «A»
- Da un' estrazione eseguita sul DWH emerge che in ogni distretto (borough) almeno l' 80% dei ristoranti riporta una valutazione «A» (in proporzione al numero di ristoranti presenti nel distretto)
- La maggior parte delle ispezioni che ha evidenziato una violazione riporta la suddetta violazione come critica. Tutte le violazioni segnalate come critiche sono potenzialmente imputabili di un' eventuale intossicazione alimentare.
- Tra i tipi di violazione più frequenti si evidenziano tra le altre «tracce di topi» e «condizioni favorevoli ai parassiti»
- Il volume delle ispezioni risulta costante negli ultimi anni fatta eccezione per l' anno 2020 che non riporta dati dopo Febbraio a causa dell' emergenza per il COVID-19. In ogni caso il volume di ispezioni è in linea con l' obbiettivo di ispezionare almeno una volta l' anno tutti i ristoranti (al momento dell' estrazione circa 27k ristoranti)
- New York risulta essere un melting pot anche dal punto di vista della proposta culinaria ma la maggior parte dei ristoranti propone una cucina «Americana»

DONE / TO DO

DONE

- Analisi
- Trasformazione e pulizia del dato
- Modellazione del dato
- Alimentazione DWH
- Creazione visualizzazione

TO DO

- Creare una pipeline di caricamento per il datawarehouse
- Modificare il notebook affinché possa essere usato per fare un refresh del DWH qualora il progetto fosse go live.

Hint per sviluppi futuri

Il programma di «restaurants grading» è in vigore in alcune città americane quali New York,Los Angeles, Milwaukee etc.
Il consumatore è al momento informato del grading solamente se arriva fisicamente davanti all' ingresso del ristorante.

Potrebbe essere interessante invece inserire tali informazioni all' interno delle schede dei ristoranti in servizi come TripAdvisor oppure di delivery quali JustEat/Uber Eats… etc

In questo modo il programma sarebbe sicuramente più efficace perché raggiungerebbe i canali che usano la maggior parte delle persone al posto del sito di OpenData della propria città.

Un primo elemento su cui riflettere potrebbe essere la difficoltà di identificare in maniera univoca il ristorante tra le piattaforme visto che l' unico elemento potrebbe essere al momento il nome del ristorante.

GRAZIE!