Choose Your Travel Experience!

BIG DATA & ANALYTICS

Web Data Ingestion, ETL & Data Quality

Milano, 03/07/2020

Chiara Teruzzi

AGENDA

OBIETTIVO

Quali sono le domande a cui si vuole rispondere con la propria soluzione

DATASET

Quali sono i dati da cui si è partiti e quali le caratteristiche

ARCHITETTURA

Quali sono state le scelte tecnologiche e di architettura per trattare il proprio Dataset

DATA QUALITY

Tecniche scelte, risultati ottenuti, problematiche riscontrate

SOLUZIONE

Soluzione Finale proposta e possibili sviluppi futuri

CONCLUSIONI

Cosa si potrebbe migliore e cosa invece è stato utile

FACTS

31 Miliardi di euro è il valore degli acquisti online effettuati dagli italiani nel 2019 (Webitmag, 2019)

10,8 miliardi di euro riguardano il mercato del turismo e dei trasporti (Webitmag, 2019)

I trasporti si confermano la categoria principale per quanto riguarda la spesa digitale (61%), seguiti da alloggi (29%) e pacchetti (10%) (Askanews, 2019)

Solo il 2% degli italiani tra i 18 e i 75 anni non ha usato internet per nessuna attività relativa alla sua ultima vacanza (Askanews, 2019)

Secondo i dati ISTAT, più della metà degli italiani (56 %) **prenota da sé i propri viaggi**, il 36,5 per cento non prenota, probabilmente perché ospite da amici o parenti, e solo il 6,6 per cento delle prenotazioni avviene tramite un'agenzia di viaggi

OBIETTIVO

La soluzione ha l'obiettivo di fornire all'utente una **Piattaforma dinamica** all'interno della quale sia possibile filtrare per **prezzo**, **giorni di viaggio**, **location** e **agenzia** le varie offerte presenti su diversi siti web.

Attraverso la piattaforma deve essere possibile anche navigare il dettaglio dell'offerta e cliccare sul **link di prenotazione** che rimanderà direttamente al sito web per proseguire l'attività.

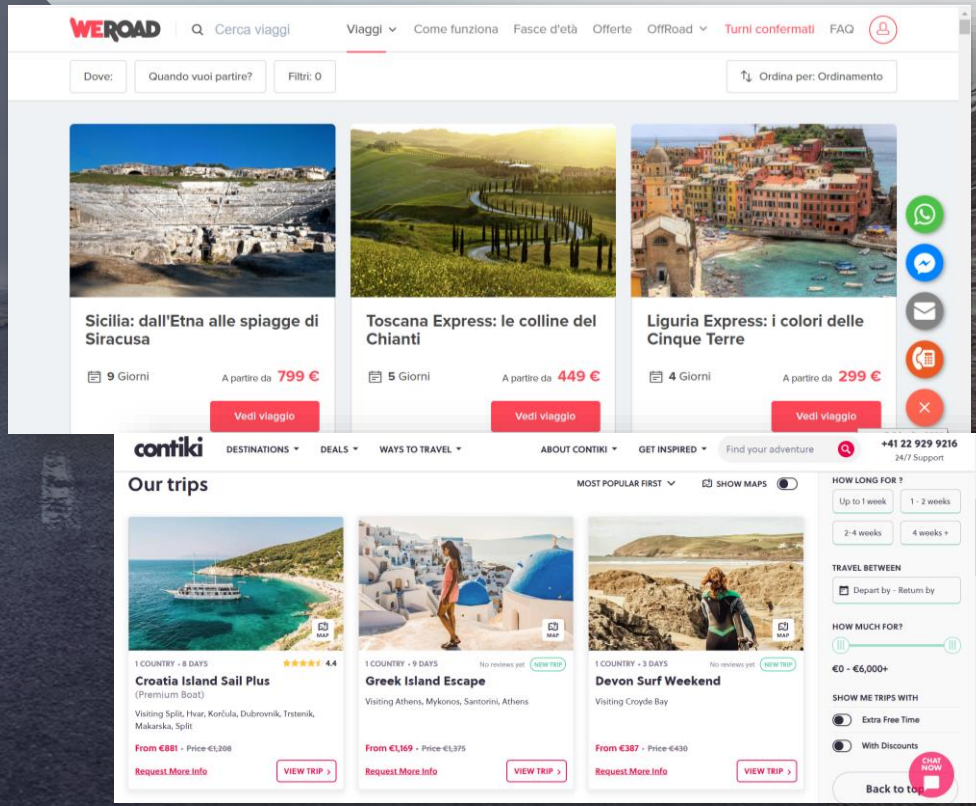
In questo modo si **riducono i tempi di ricerca** e **prenotazione agevolando l'esperienza dell'utente**.

DATASET – Overview

Il Dataset è costituito dai dati relativi alle offerte di viaggio delle agenzie WeRoad e Contiki, specializzate in **viaggi di gruppo organizzati**.
I dati sono stati scaricati dai rispettivi siti web tramite scraping in Python e caricati come .csv in Power BI.

Le informazioni che sono state raccolte fanno riferimento a:

- **Titolo Offerta**
- **Prezzo (in €)**
- **Durata del viaggio**
- **URL di dettaglio**



DATASET – Big Data

I Dati analizzati non possono essere esattamente classificati come **Big Data**, in quanto non rispondono a tutte le 4V che li caratterizzano.

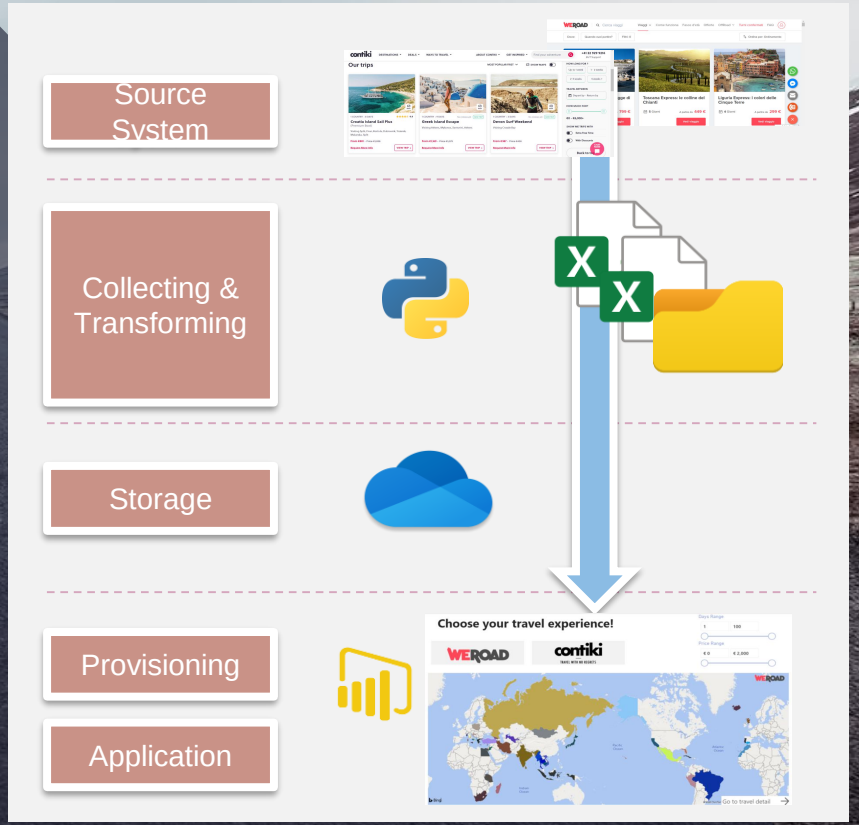
Di seguito comunque si riportano le caratteristiche del Dataset:

- **VOLUME:** il volume dei dati è limitato in quanto il Dataset potrà includere più siti, ma il numero di offerte medio sarà sempre tra i 200 e 300 viaggi (il format è viaggio organizzato)
- **VARIETY:** i dati saranno relativi alla stessa tematica ma avranno strutture e formattazioni differenti a seconda del sito web
- **VELOCITY:** le offerte viaggio cambieranno / si aggiorneranno stagionalmente ma non ci sarà necessità di aggiornare con una frequenza elevata il dataset
- **VERACITY:** il dato è veritiero ma richiederà una fase di Qualità prima di presentarlo all'utente finale

ARCHITETTURA - Attuale

I **dati** vengono dai 2 siti web WeRoad e Contiki.
La parte di **ETL** è stata svolta tramite script **Python** che ha permesso di estrarre i dati del sito web (in **HTML**) e caricarli su un file di tipo **csv**.
I dati sono stati poi inseriti in una **cartella** e caricati su **OneDrive**.

Il Semantic Layer e quindi tutta la parte di modellazione è stata svolta tramite **Power BI**.
Prima di presentare il dato si è effettuata una parte di Data Quality tramite **PowerQuery (M Query)**.

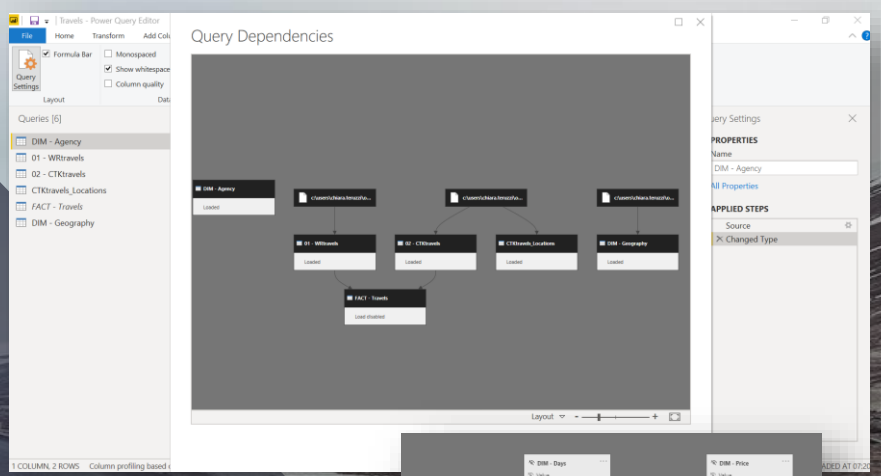


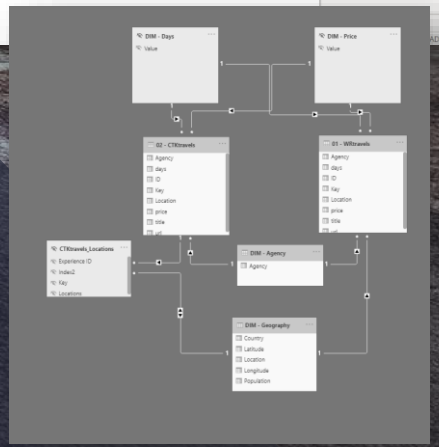
DATA QUALITY

La «Data Quality» del Dataset è stata svolta principalmente tramite Power BI (M Query).

Queste le maggiori criticità:

- **La location** è stata estratta dai titoli/sottotitoli dell'offerta
- Le location tra i due dataset **non erano omogenee** (lingua e granularità differenti)
- Per Contiki, in **un unico campo erano presenti più destinazioni** a parità di viaggio (separate da «,»)
- I due siti mostrano le stesse info di viaggio ma con **strutture differenti**
- **A seconda nella località da cui si naviga il sito si possono avere a disposizione informazioni differenti (es: prezzo)**





DATA QUALITY - Dimension

Di seguito riporto le principali dimensioni di Qualità applicate al Dataset in oggetto:

- Accuratezza → i dati risultano accurati in quanto tutte le informazioni presenti sul dito sono state caricate così come sono.
- Completezza → per i due siti i dati sono completi, per avere una visione completa delle offerte viaggio andrebbero estratti i dati da ulteriori siti
- Volatilità → i dati non cambiano con una frequenza elevata ma stagionalmente (mensilmente)
- Aggiornamento → il dato ad oggi si aggiorna manualmente, si potrebbe però schedulare l'aggiornamento mensilmente
- Tempestività → il dato viene aggiornato in tempo per i fini di progetto (va considerata la frequenza delle offerte proposte ma che dovrebbe essere relativa al mese)
- Consistenza → i dati risultano consistenti

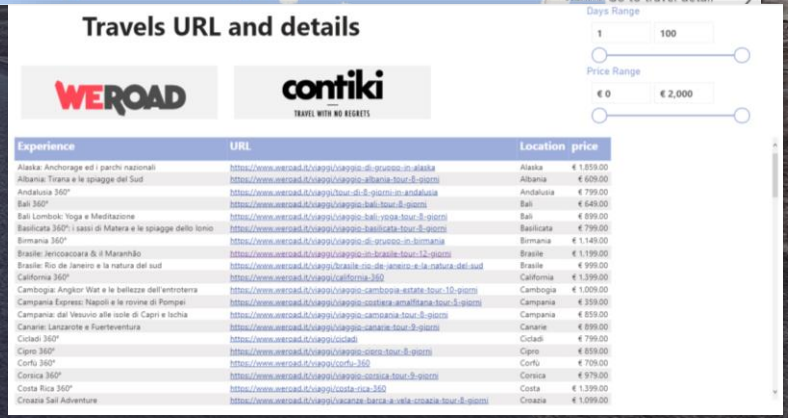
SOLUZIONE

La soluzione è quindi costituita da un **report di Power BI**, con una pagina navigabile dinamicamente ed una pagina di dettaglio.

Il report può essere pubblicato su **Power BI Service e condiviso tramite link**.

I dati, una volta gestiti più siti, possono essere gestiti da un **R-DBMS** in quanto le informazioni che vengono storicizzate saranno sempre le stesse.

Lo **script Python** dovrà invece essere aggiornato per scaricare sempre le stesse informazioni dai siti web.



CONCLUSIONI

- La parte più complessa nella raccolta dati è la **Data Quality**
- Trattare dati da **fonti dati diverse** implica la necessità di rielaborare tutte le informazioni raccolte in modo che siano **confrontabili tra di loro**
- Bisogna sempre cercare di comprendere la vera necessità di Business e per ogni necessità utilizzare lo **strumento corretto**
- **La soluzione più complessa non è sempre la soluzione corretta**

THANKS!

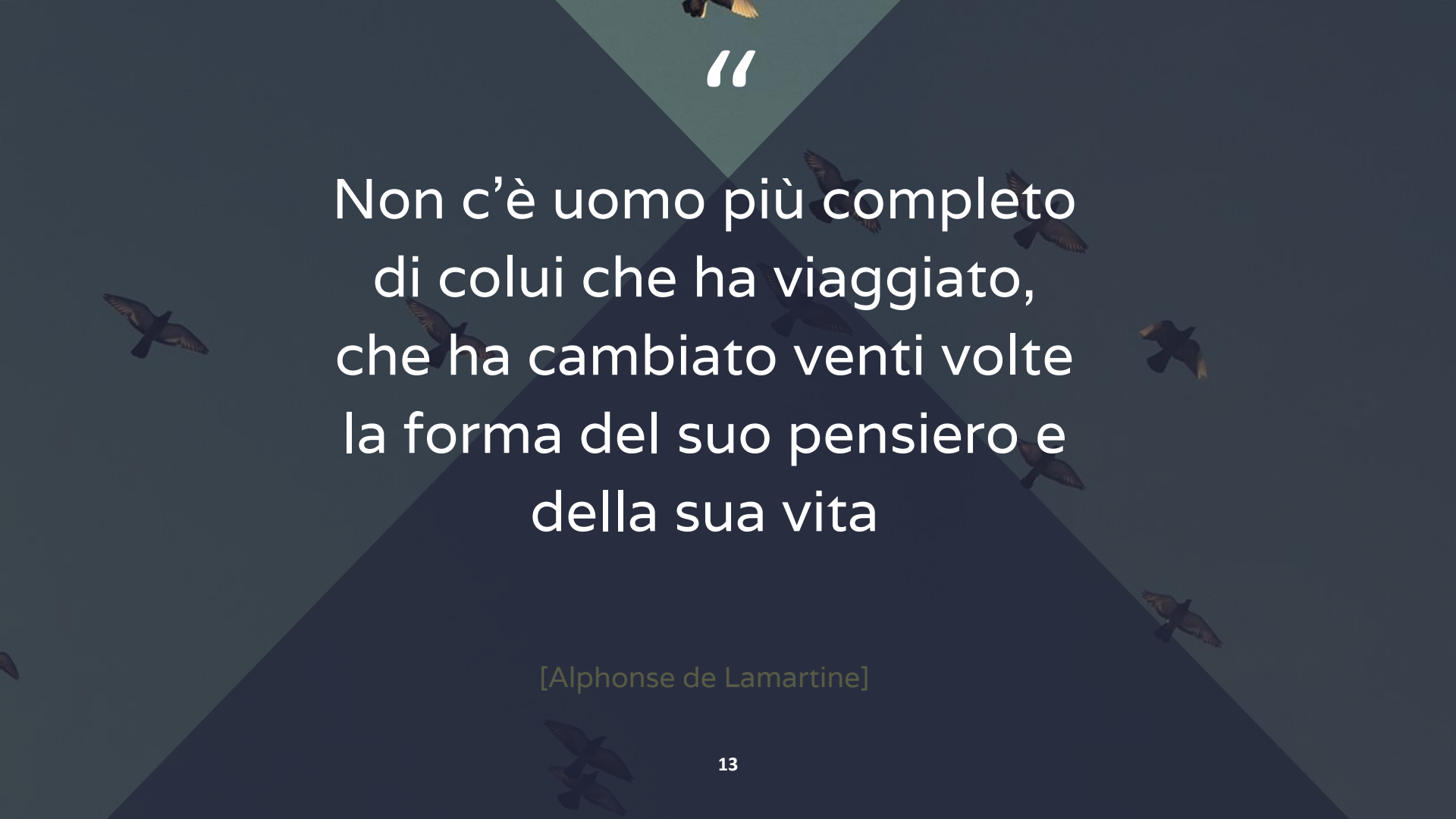
Any questions?

You can find me at:


Chiara.teruzzi@outlook.it


https://github.com/ChiaraTeruzzi


https://www.linkedin.com/in/chiara-teruzzi-040024b4/



“

Non c’è uomo più completo
di colui che ha viaggiato,
che ha cambiato venti volte
la forma del suo pensiero e
della sua vita

[Alphonse de Lamartine]