I

Marvel Graph DB

DATA MANAGEMENT

II

Data Management

Progetto AA 2021/22

Fabrizio Cominetti, Davide Abete, Ruben Agazzi

III

Introduzione

Il progetto ha avuto origine dall'idea di realizzare un ampio database relativo al mondo **Marvel**, una base di dati che permettesse di navigare al suo interno tra i vari prodotti Marvel, fra cui personaggi, film, serie tv e fumetti.
L'obiettivo del progetto è di ottenere una sorta di "**rete sociale**" di super-eroi o personaggi Marvel, i quali sono collegati ai rispettivi fumetti, film, serie tv in cui sono presenti, oppure ancora presentano i rispettivi collegamenti interpersonali fra personaggi ed altri personaggi. Per questi motivi la scelta in fase di costruzione del progetto è ricaduta su un modello a **grafo**. I dati sono stati ottenuti mediante l'utilizzo di **API** e web **scraping**, per arrivare alla costruzione del database tramite l'utilizzo della piattaforma open-source **Neo4j**.

Web API

Utilizzo API ufficiale fornita dalla Marvel [Marvel Developer Portal]

1. Ottenimento dei dati dei personaggi in formato JSON
2. Salvataggio su file CSV dei dati ottenuti
3. Ottenimento dei dati dei fumetti in formato JSON
4. Salvataggio su file CSV dei dati ottenuti

Web Scraping

Utilizzo della Marvel Cinematic Universe Wiki [Marvel Cinematic Universe Wiki | Fandom]

1. Ottenimento della pagina con lista dei nomi dei personaggi con i relativi link alla pagina personale
2. Per ogni personaggio, acquisizione delle informazioni rilevanti
3. Per ogni film trovato nelle pagine dei personaggi, selezionare la pagina relativa e recupero delle informazioni del film
4. Per ogni serie, trovata nelle pagine dei personaggi, ottenere le informazioni della serie in questione
5. Salvataggio temporaneo all’interno di file CSV

V

Data Cleaning Biografia

Principali problematiche:

1. Il notebook salva i vari paragrafi delle biografie come una lista di stringhe
2. Il notebook salvava alcuni paragrafi della biografia più volte, quindi è necessario filtrarli
3. Infine, viene fatto un escaping dei caratteri speciali, come ad esempio il carattere "\n"

Data Cleaning Relazioni

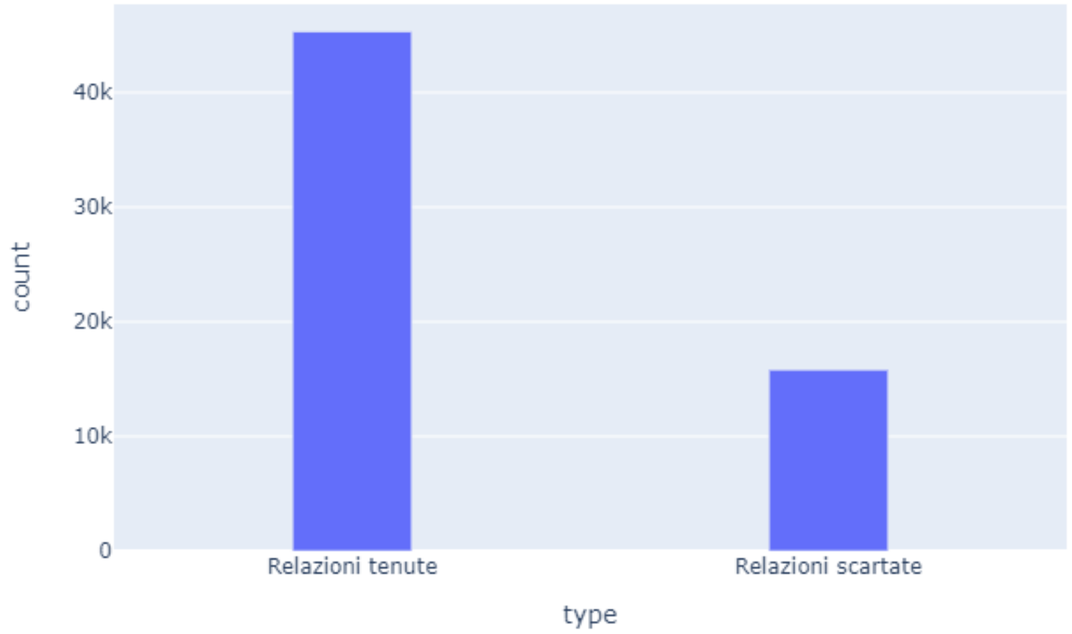
La principale problematica riguarda l'assenza in alcune relazioni del nome del personaggio interessato. Per risolvere questo problema il notebook esegue le seguenti operazioni:

1. Per ogni personaggio vengono recuperate le relative relazioni
2. Per ogni relazione ottenuta viene effettuato un controllo sul nome del personaggio della relazione, se il nome non è presente all'interno della lista di personaggi la relazione viene scartata

VI

Data Cleaning
Relazioni

Numero di relazioni tenute (sx) e relazioni scartate (dx)



Completezza

Sono state effettuate altre analisi riguardanti la completezza dei dati, in particolare riguardo alla completezza a livello di tabella dei vari dati ottenuti. Per calcolare la completezza abbiamo utilizzato la seguente formula: (NumeroMissingValuesTabella) / (NumeroColonne ∗ NumeroRighe)

Personaggi Web API

Completezza dati relativi ai personaggi ottenuti tramite consultazione della web API pari al 92, 6%

Fumetti Web API

Completezza dati relativi ai fumetti ottenuti tramite Web API pari al 83, 9%

Film Web Scraping

Completezza dati relativi ai Film ottenuti tramite Web Scraping pari al 89, 1%

Personaggi Web Scraping

Completezza dati relativi ai personaggi ottenuti tramite Web Scraping pari al 95, 6%

Serie TV Web Scraping

Completezza dati relativi alle serie TV ottenuti tramite Web Scraping pari al 96, 7%

Ridondanza

Un’ulteriore analisi di qualità effettuata consiste nell’analisi di ridondanza dei valori a livello di tabella, in particolare è stato contato, per ogni tabella, il numero di righe con attributo identificativo, come ad esempio nome o titolo, uguali.

Personaggi Web API

1 personaggio duplicato

Personaggi Web Scraping

0 personaggi duplicati

Fumetti Web API

2216 fumetti duplicati

Serie TV Web Scraping

0 serie TV duplicate

Film Web Scraping

0 film duplicati

Schema Transformation

È stato effettuato un processo di 'Reverse Engineering' per capire cosa rappresentassero i dati della web API. Tale operazione non è stata fatta per lo scraping, in quanto lo schema era già chiaro a priori.

Corrispondences Investigation

Si è notato che si possono unire le due basi di dati sulla base dei personaggi, è possibile stabilire un collegamento fra i personaggi sulla base del loro nome. Un personaggio è identificato tramite un nome privo di parentesi e un'eventuale variante del personaggio è identificata da un nome all'interno della parentesi.

Schema Integration

Questa fase avviene durante l'inserimento dei dati nel database: viene creato all'interno del database un nodo di tipo character se non è già presente un nodo dello stesso tipo con lo stesso nome.

X

Struttura DB
Nodi

I nodi presenti all'interno del database sono:

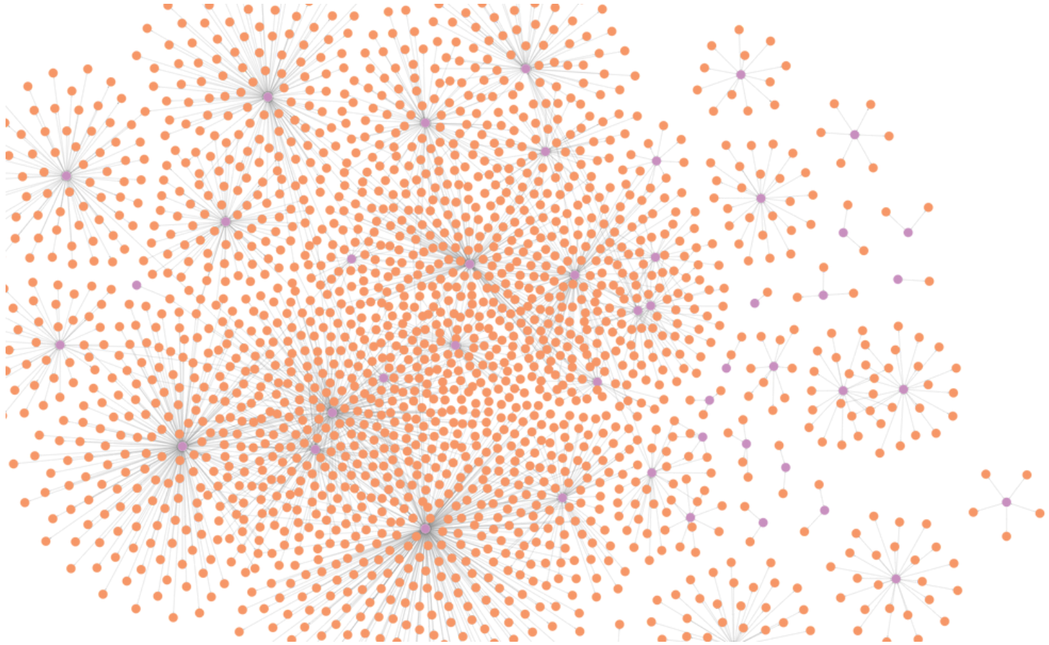
1. **Comic:** fumetto Marvel [ID, codici di identificazione, descrizione se presente, formato e numero di pagine]
2. **Movie:** film Marvel [trama, registi, scrittori, produttori, compositori, incassi, durata e data di rilascio]
3. **Tv Show:** serie tv Marvel [trama della prima stagione, registi, showrunner, produttori e compositori]
4. **Character:** versione generica di un personaggio [nome del personaggio]
5. **Character Variant:** variante precisa di un personaggio [nome, descrizione e biografia]

Struttura DB
Relazioni

Le relazioni presenti fra i nodi all'interno del database sono:

1. **In Film:** relazione che collega un nodo «Character Variant» ad un nodo film
2. **In Serie:** relazione che collega un nodo «Character Variant» ad un nodo Tv Show
3. **In fumetto:** relazione che collega un nodo «Character Variant» ad un nodo «Comic»
4. **Conosce:** relazione che collega un nodo «Character Variant» ad un nodo «Character Variant»
5. **Variante di:** relazione che collega un nodo di tipo «character» ad un nodo di tipo «character variant»

Statistiche DB



Nodi

All'interno del database sono stati creati in totale 60411 nodi così distribuiti:

Etichetta Nodo	Numero Di Nodi
Character	4609
Character Variant	5113
Comic	50607
Movie	43
Tv Show	19

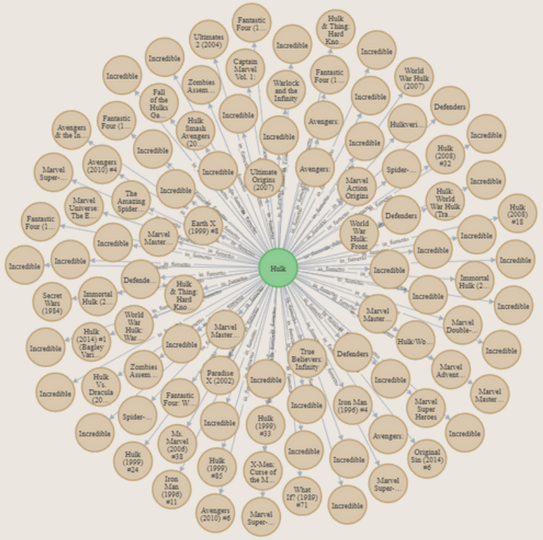
Relazioni

Fra i nodi del database sono state create 117763 relazioni distribuite come di seguito:

Etichetta Relazione	Numero Di Relazioni
Conosce	19040
In Film	531
In Serie	2405
In Fumetto	90503
Variante di	5284

Query 1

MATCH (i:character_variant)-[r:conosce]->
(c:character_variant)-[p:in_film]
->(m:movie) WHERE i.name = "Iron Man" AND m.title
= "The Avengers"
RETURN i,r,c,p,m



Query 2

MATCH (c:character_variant)-[r:in_fumetto]->
(f:comic) WHERE c.name =
"Hulk" RETURN c,f,r LIMIT 100

XIII

FINE

Fabrizio Cominetti, Davide Abete, Ruben Agazzi