GESTIONE AVANZATA DEI DATI 2015/2016

Giovanni De Costanzo *Giuseppe Angri*

SPECIFICA DEL PROBLEMA - 1

- Cerchi un tutorial che ti aiuti ad apprendere un nuovo linguaggio di programmazione, un suo framework o una particolare tecnologia, ma non sai dove cercare e magari non vorresti spendere soldi.
- Preferisci studiare dai libri, quindi vuoi sapere quali trattano quell'argomento e visualizzare il miglior prezzo d'acquisto.
- Hai bisogno di una documentazione ufficiale.
- Ti piacerebbe partecipare a qualche evento informatico, o cercare una community con i tuoi stessi interessi.

Problema: per ricercare tutte queste informazioni è necessario visitare più siti -> ricerca noiosa e perdita di tempo

SPECIFICA DEL PROBLEMA – 2

DevHelper è un motore di ricerca che velocizza tutto, raccogliendo per te queste informazioni.

SPECIFICA DEL PROBLEMA – 3

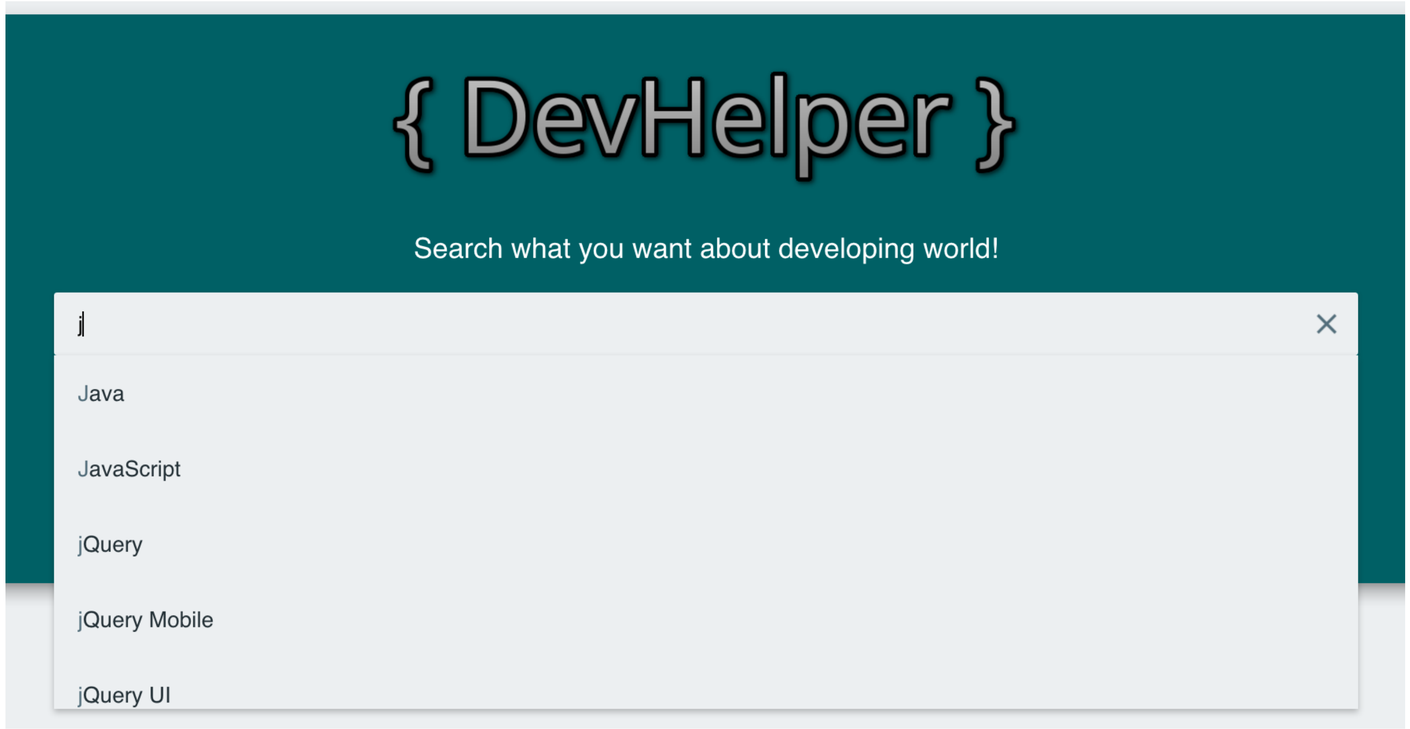
- Cerca da più fonti tutorial e videocorsi.
- Cerca i libri di tuo interesse, mostrandoti i prezzi da differenti store online.
- Cerca tutti i prossimi eventi e le communities relative all’argomento di tuo interesse oppure nella tua città.
- Se sei interessato ad una documentazione, ti indirizza direttamente alla pagina di interesse.
- Se cerchi un linguaggio di programmazione, ti permette di vedere anche quali sono i framework associati.

DEMO

MODALITÀ DI UTILIZZO

- Ricerca per Keyword inserendo:
  - Linguaggio di Programmazione
  - Framework
  - Tecnologia

- Autocompletamento con aggiornamento dinamico dei suggerimenti



FONTI - 1

- **Udemy :**
  - Utilizzata per ottenere i videotutorial
  - Web Scraping

- **Lynda.com :**
  - Utilizzata per ottenere i videotutorial
  - Scraping by import.io

- **Microsoft Virtual Academy :**
  - Utilizzata per ottenere i videotutorial
  - API

FONTI - 2

- **ISBN.Directory** :
  - Utilizzato per la ricerca dei libri e delle loro informazioni
  - API

- **Amazon** :
  - Utilizzato per la ricerca dei prezzi eBook e libri cartacei
  - Scraping by import.io

- **Libreria Universitaria** :
  - Utilizzato per la ricerca dei dei prezzi dei libri cartacei
  - Web Scraping

FONTI - 3

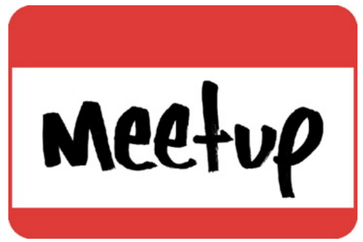
- **TutorialsPoint :**
  - Utilizzato per la ricerca di tutorial in lingua inglese
  - Web Scraping

- **HTML.IT :**
  - Utilizzato per la ricerca di tutorial in lingua italiana
  - Web Scraping

FONTI - 4

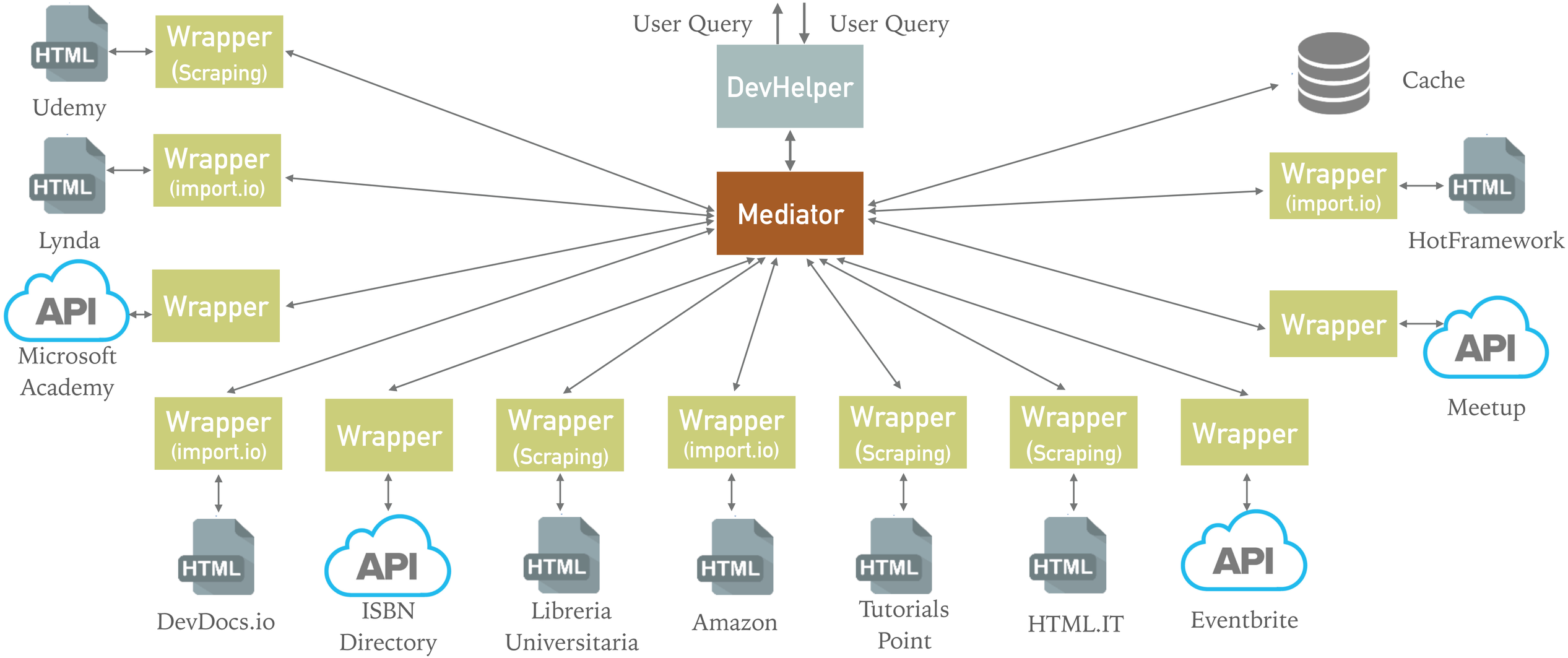
- **Eventbrite :**
  - Utilizzato per la ricerca di eventi
  - API

- **Meetup :**
  - Utilizzato per la ricerca di gruppi/community
  - API



- **HotFrameworks.com :**
  - Utilizzato per la ricerca di web frameworks
  - Scraping by import.io

ARCHITETTURA DEL SISTEMA - (12 FONTI)



CACHE

Per ogni query effettuata, viene salvato il risultato in cache.

Ad ogni query salvata in cache, viene associato un tempo di scadenza differente, in base alla volatilità della fonte.

- Videotutorial: 3 giorni
- Libri: 15 giorni
- Prezzi libri: 1 giorno
- Tutorial: 3 giorni
- Eventi: 1 giorno
- Gruppo: 3 giorno
- Framework: 15 giorni

QUERYING SUL SISTEMA

- Le query sul sistema vengono effettuate direttamente sugli schemi globali. Ognuno di essi coincide con uno script PHP, il quale riceve in input la keyword inserita dall'utente.
- Lo script interroga le sorgenti (tramite scraping o api) oppure prende il contenuto salvato nel Database di cache.
- Possiamo distinguere le sorgenti utilizzate in:

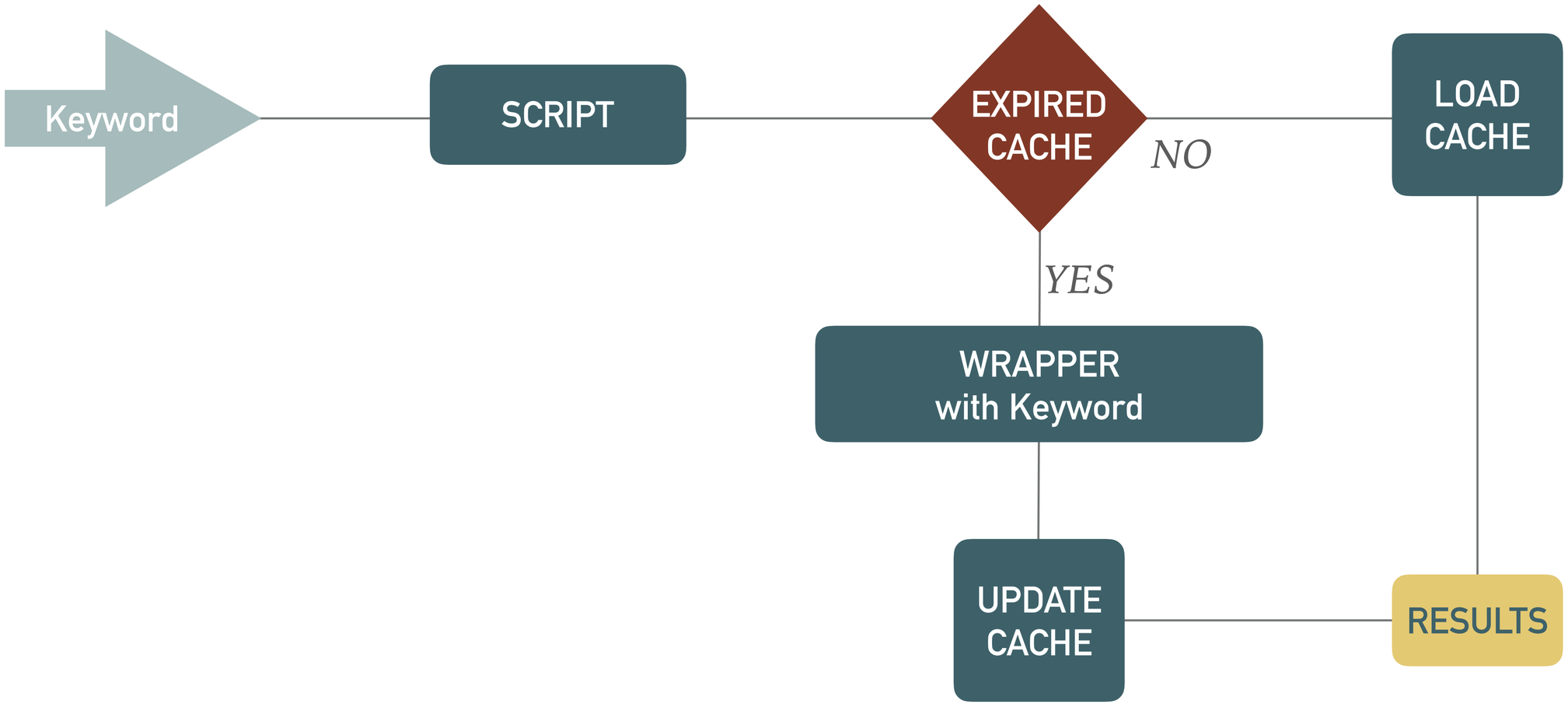
1. Sorgenti che permettono la ricerca per Keyword

   oppure

2. Sorgenti che restituiscono tutti i contenuti presenti

QUERYING SUL SISTEMA – CASO 1

- Nel primo caso, gli script funzionano secondo il seguente diagramma:

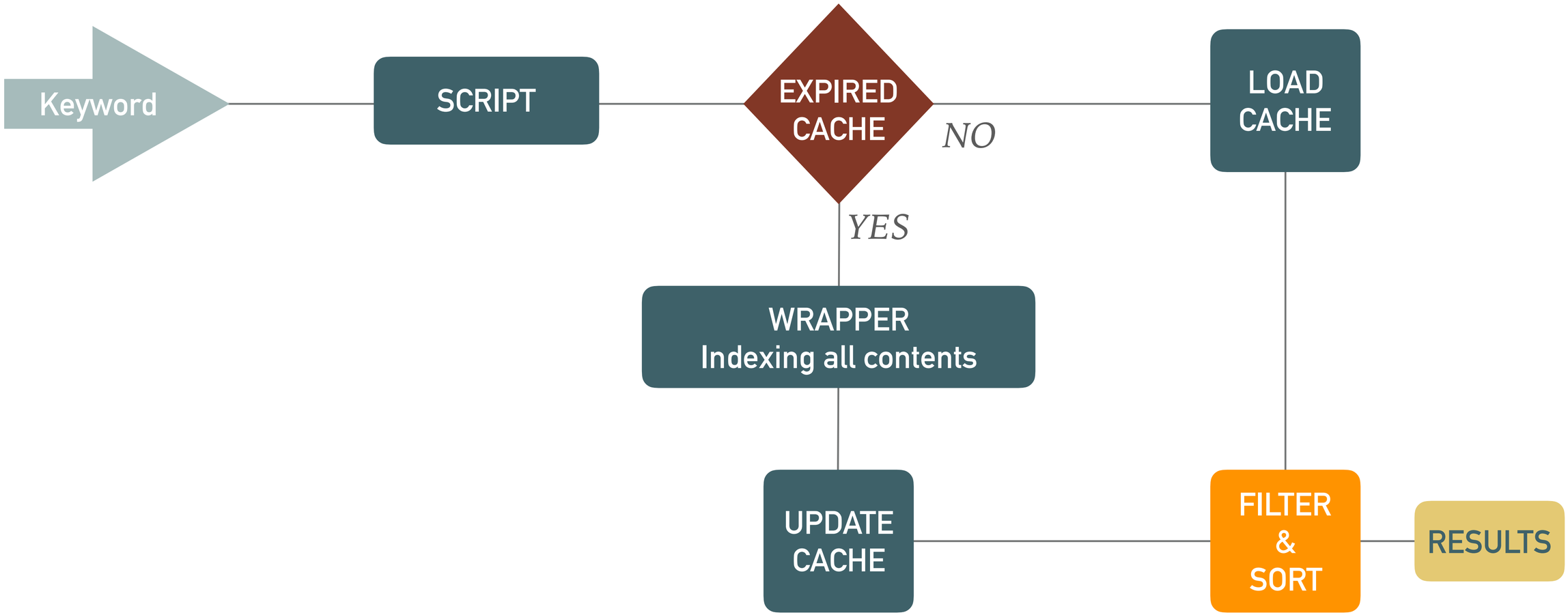


QUERYING SUL SISTEMA – CASO 1

- Lo schema descritto viene utilizzato per le seguenti fonti:
  - Udemy
  - Lynda
  - Microsoft Academy
  - ISBN directory
  - Amazon
  - Libreria Universitaria
  - Eventbrite
  - Meetup

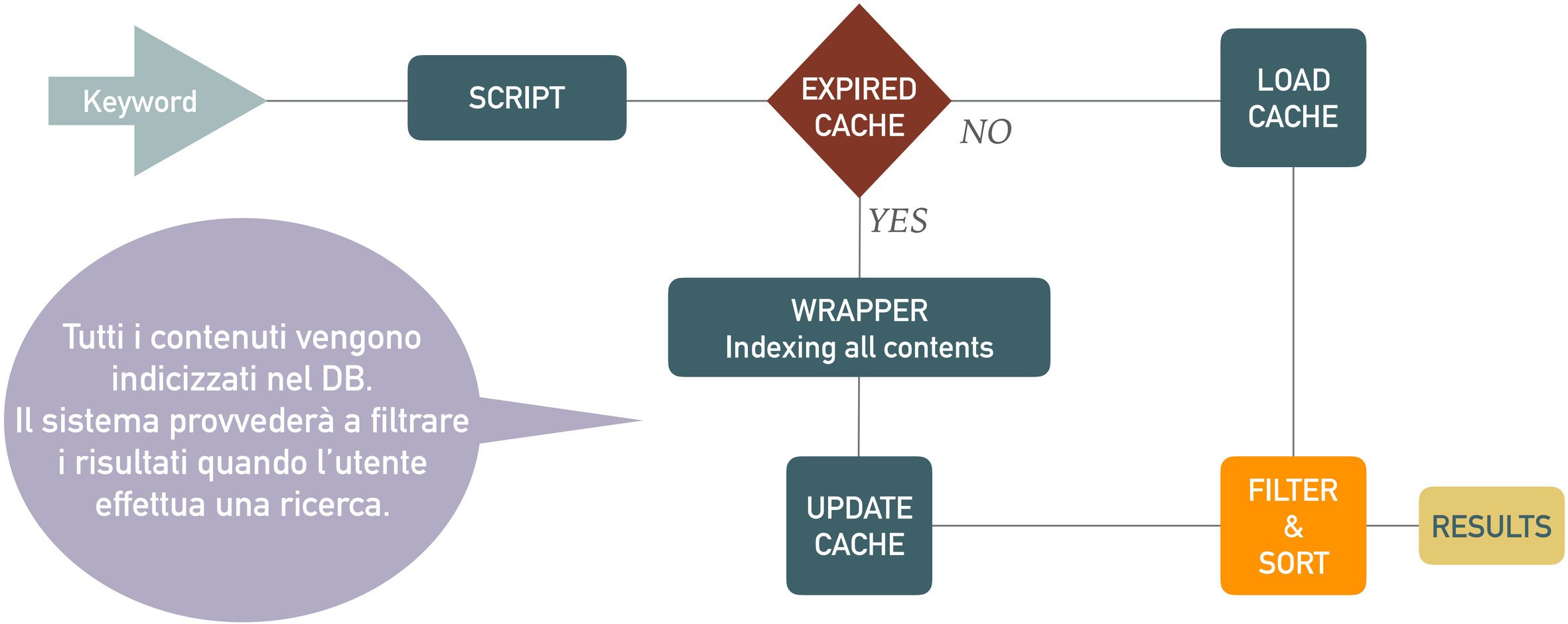
QUERYING SUL SISTEMA – CASO 2

- Nel secondo caso, gli script funzionano secondo il seguente diagramma:



QUERYING SUL SISTEMA – CASO 2

- Nel secondo caso, gli script funzionano secondo il seguente diagramma:



QUERYING SUL SISTEMA – CASO 2

- Lo schema descritto viene utilizzato per le seguenti fonti:
  - HTML.IT
  - TutorialsPoint
  - HotFrameworks.com

COLLEGAMENTO FONTI (JOIN) - 1

- Sono state implementate le seguenti Join:
  - Ricerca dei Prezzi per Libro
  - Ricerca dei Libri per Autore di un Videocorso
  - Ricerca di una Community per la città di un Evento
  - Ricerca di un Evento per la città di una Community

COLLEGAMENTO FONTI (JOIN) – 2

- La ricerca dei libri viene effettuata su ISBN.Directory
  - Per ogni libro restituito utilizziamo il codice ISBN per ricercare i prezzi dei libri cartacei su Amazon e Libreria Universitaria, e i prezzi degli eBook da Amazon.

- La ricerca dei videotutorial viene effettuata su Udemy, Lynda e M.V.A.
  - Per ogni risultato utilizziamo le informazioni sull'autore per ricercare eventuali libri di testo scritti dallo stesso, su ISBN.Directory.

- Gli eventi vengono ricercati su Eventbrite, mentre le communities su Meetup
  - Per ogni community vengono ricercati gli eventi tenuti nella stessa città
  - Per ogni evento vengono ricercate le communities presenti nella stessa città

COLLEGAMENTO FONTI (JOIN) – 3

- Per mantenere alte le performance del sistema, nel caso in cui il numero dei risultati ottenuti è elevato, le query di join vengono effettuate solamente su richiesta.

  Es.: da una query di ricerca si ottengono *n* risultati.
  Se le join fossero sempre eseguite, sarebbero necessarie altre *n* query.
  Quindi andrebbero effettuate *n+1* query prima di mostrare i risultati.

- Dunque, il sistema restituisce prima i risultati di una ricerca e l'utente, se interessato, richiede le ulteriori informazioni previste dalle join descritte in precedenza.

DESCRIZIONE SCHEMI LOCALI - 1

- **Udemy** (url, title, author, price, image)
- **Lynda** (url, title, author, image, description)
- **MicrosoftAcademy** (url, title, author, image, description)
- **IsbnDirectory** (title, description, author, isbn, img, publication_date, page_number)
- **LibreriaUniversitaria** (isbn, url, price)
- **Amazon** (isbn, format, url, price)
- **Eventbrite**
  - **event** (name, description, url, date_start, date_end, image, venue_id)
  - **venue** (id, city, region)

DESCRIZIONE SCHEMI LOCALI - 2

- **TutorialsPoint** (title, category, url)
- **HTML.IT** (title, category, url, keywords)
- **Meetup** (url, title, description, city, region, members_num, img)
- **HotFrameworks**
  - **lang_frameworks** (language, framework)
  - **framework_details** (framework, description, official_page_url)

WRAPPER - 1

- **Udemy -** https://www.udemy.com/courses/search/?lr=1&q= + <search_string>
  - url -> `//*[@id="courses"]/li/a/@href`
  - title -> `//*[@id="courses"]/li/div/div/div[@class="title-wr"]/span`
  - author -> `//*[@id="courses"]/li/div/div/span[@class="ins"]`
  - price -> `//*[@id="courses"]/li/div[2]/div[4]/span`
  - image -> `//*[@id="courses"]/li/div/span/img/@src`

WRAPPER – 2

- **LibreriaUniversitaria -** http://www.libreriauniversitaria.it/ricerca/query/+<isbn>
  - url -> `//*[@id="colmain"]/div[4]/div/div[@class="search-image"]/a/@href`
  - price -> `//*[@id="colmain"]/div[4]/div/div[@class="search-details-ricerca"]/div[@class="info-prezzosconto"]/div[@class="product_our_price"]`
- **HTML.IT -** http://www.html.it/development/guide/
  - url -> `//dd/h3/a/@href`
  - title -> `//dd/h3/a`
  - category -> `//dd/ul/li/a`
  - keywords -> `//dd/a`

WRAPPER – 3

- **TutorialsPoint :** http://www.tutorialspoint.com/tutorialslibrary.htm
  - category -> `//ul[@id='$id']/preceding-sibling::h4`
  - title -> `//ul[@id='$id']/li/a`
  - url -> `//ul[@id='$id']/li/a/@href`

- **Lynda (import.io API)**

https://api.import.io/store/connector/12513729-2036-425e-9600-1325a160d3e7/_query?&_apikey=<API_KEY>&input=webpage/url:http%3A%2F%2Fwww.lynda.com%2Fsearch%3Fq%3D + <SEARCH_STRING>

WRAPPER - 4

- **Amazon e HotFrameworks.com (import.io API)**

  API calls molto simili a quella vista per Lynda

- **Eventbrite (API)**

  Event: https://www.eventbriteapi.com/v3/events/search/?venue.country=IT&token=<DEV_TOKEN>&format=json&categories=102&q=<SEARCH_STRING>

  Venue: https://www.eventbriteapi.com/v3/venues/<VENUE_ID>/?token=<DEV_TOKEN>&format=json

- **Meetup (API)**

  https://api.meetup.com/find/groups?key=<API_KEY>&sign=true&photo-host=public&country=IT&text=<SEARCH_STRING>

WRAPPER – 5

- **Microsoft Virtual Academy (API)**

  http://api-mlxprod.microsoft.com/sdk/search/v1.0/5/courses

  API call e parametri in formato json tramite POST

- **ISBN.Directory (API)**

  http://isbn.directory/ajax?ajax=search&q=<SEARCH_STRING>&page=<PAGE_NUMBER>

SCHEMA GLOBALE

- **Videocorsi** (url, title, author, price, image, description)
- **Books** (title, description, author, isbn, img, publication_date, page_number)
- **BookPrice** (isbn, price, url, store, format)
- **Eventi** (name, description, url, date_start, date_end, image, city, region)
- **Frameworks** (language, framework, description, official_page_url)
- **Tutorial** (title, category, url, lang)
- **Community** (url, name, description, city, region, members_num, img)

MAPPING GAV - 1

- **Videocorsi** (url, title, author, price, image, description) :-
  Udemy (url, title, author, price, image) ^ description = ‘’

- **Videocorsi** (url, title, author, price, image, description) :-
  Lynda (url, title, author, image, description) ^ price = ‘Free’

- **Videocorsi** (url, title, author, price, image, description) :-
  MicrosoftAcademy (url, title, author, image, description) ^ price = ‘Free’

MAPPING GAV – 2

- **Books** (title, description, author, isbn, img, publication_date, page_number) :-
  IsbnDirectory (title, description, author, isbn, img, publication_date, page_number)
- **BookPrice** (isbn, price, url, store, format) :-
  IsbnDirectory (_ , _ , _ , isbn, _ , _ , _) ^
  LibreriaUniversitaria (isbn, url, price) ^ store = ‘LibreriaUniversitaria’
  ^ format = ‘Copertina Flessibile’
- **BookPrice** (isbn, price, url, store, format) :-
  IsbnDirectory (_ , _ , _ , isbn, _ , _ , _) ^
  Amazon (isbn, format, url, price) ^ store = ‘Amazon’

MAPPING GAV – 3

- **Tutorial** (title, category, url, lang) :-
  TutorialsPoint (title, category, url) ^ lang = ‘ENG’

- **Tutorial** (title, category, url, lang) :-
  HTML.IT (title, category, url, _) ^ lang = ‘IT’

- **Community** (url, name, description, city, region, members_num, img) :-
  Meetup (url, name, description, city, region, members_num, img)

MAPPING GAV - 4

- **Eventi** (name, description, url, date_start, date_end, image, city, region) :-
  Eventbrite.event (name, description, url, date_start, date_end, image, venue_id) ^
  Eventbrite.venue (venue_id, city, region)

- **Frameworks** (language, framework, description, official_page_url) :-
  HotFrameworks.lang_frameworks (language, framework) ^
  HotFrameworks.framework_details (framework, description, official_page_url)

MAPPING LAV - 1

- **Udemy** (url, title, author, price, image) :-
  Videocorsi (url, title, author, price, image, _)

- **Lynda** (url, title, author, image, description) :-
  Videocorsi (url, title, author, _ , image, description)

- **MicrosoftAcademy** (url, title, author, image, description) :-
  Videocorsi (url, title, author, _ , image, description)

MAPPING LAV – 2

- **IsbnDirectory** (title, description, author, isbn, img, publication_date, page_number) :-
  Books (title, description, author, isbn, img, publication_date, page_number)

- **LibreriaUniversitaria** (isbn, url, price) :-
  BookPrice(isbn, price, url, store, _) ^ store = ‘LibreriaUniversitaria’

- **Amazon** (isbn, format, url, price) :-
  BookPrice(isbn, price, url, store, format) ^ store = ‘Amazon’

MAPPING LAV - 2

- **Eventbrite.event** (name, description, url, date_start, date_end, image, venue_id) :-
  Eventi (name, description, url, date_start, date_end, image, _ , _)

- **Eventbrite.venue** (id, city, region) :-
  Eventi (_ , _ , _ , _ , _ , _ , city, region)

- **HotFrameworks.lang_frameworks** (language, framework,) :-
  Frameworks (language, framework, _ , _)

- **HotFrameworks.framework_details** (framework, description, official_page_url) :-
- **Frameworks** (_ , framework, description, official_page_url)

MAPPING LAV – 3

- **TutorialsPoint** (title, category, url) :-
  Tutorial (title, category, url, _)

- **HTML.IT** (title, category, url, keywords) :-
  Tutorial (title, category, url, _)

- **Meetup** (url, title, description, city, region, members_num, img) :-
  Community (url, name, description, city, region, members_num, img)

QUERY - 1

Ricerca dei libri scritti dallo stesso autore di un videotutorial

- **BooksByVideotutorialAuthor** (author, title, description, isbn, img, publication_date, page_number) :-
  Videocorsi (_, _ , author, _ , _ , _) ^
  Books (title, description, author, isbn, img, publication_date, page_number)

- SELECT b.author, b.title, b.description, b.isbn, b.img, b.publication_date, b.page_number
  FROM Books AS b, Videocorsi AS v
  WHERE b.author = v.author AND b.author = ‘Ben Tristem’

QUERY - 2

A partire dai libri trovati da una ricerca, recupera i prezzi delle versioni cartacee ed eBook.

- **BookPrices** (isbn, price, store, url, format) :-
  Books (_ , _ , _ , isbn, _ , _ , _) ^
  BookPrice(isbn, price, url , store, format)

- SELECT bp.isbn, bp.price, bp.store, bp.url
  FROM BookPrice AS bp, Books AS b
  WHERE bp.isbn = b.isbn
  AND b.title = ‘JavaScript - The definitive guide’

QUERY - 3

Cerca tutti gli eventi in una data città

- **eventiPerCitta** (city, region, name, description, url, date_start, date_end, image) :-
  Eventi (name, description, url, date_start, date_end, image, city, region)

- SELECT city, name, description, url, date_start, date_end, image
  FROM Eventi
  WHERE city = 'Salerno'

QUERY - 4

Cerca tutte le communities in una data città

- **communitiesPerCitta** (city, region, url, name, description, members_num, img) : - **Community** (url, name, description, city, region, members_num, img)

- SELECT city, region, url, name, description, members_num, img
  FROM Community
  WHERE city = ‘Salerno’

QUERY - 5

Cerca tutte le communities nella stessa città di un evento

- **communitiesCittaEvento**(city, region, url, title, description, members_num, img) :-
  Eventi (_ , _ , _ , _ , _ , _ , city, _),
  Community (url, title, description, city, region, members_num, img)

- SELECT c.url, c.title, c.description, c.city, c.region, c.members_num, c.img
  FROM Community AS c, Eventi AS e
  WHERE c.city = e.city
  AND e.name= ‘Codemotion Roma 2016’

QUERY - 6

Cerca tutti gli eventi nella stessa città di una community

- **eventiCittaCommunity** (city, region, name, description, url, date_start, date_end, image) :-
  Eventi (name, description, url, date_start, date_end, image, city, region)
  Community (_, _ , _ , city, _ , _ , _)

- SELECT e.city, e.region, e.name, e.description, e.url, e.date_start, e.date_end, e.image
  FROM Eventi AS e, Community AS c
  WHERE e.city = c.city
  AND c.name = ‘JavaScript & NodeJS community Salerno’

GAV QUERY UNFOLDING - 1

➤**BooksByVideotutorialAuthor** (author, title, descript, isbn, img, publication_date, page_number) :-
Books (title, descript, author, isbn, img, publication_date, page_number) ^
Videocorsi (_, _ , author, _ , _ , _)

➤**BooksByVideotutorialAuthor** (author, title, descript, isbn, img, publication_date, page_number) :-
IsbnDirectory (title, descript, author, isbn, img, publication_date, page_number) ^
Videocorsi (_, _ , author, _ , _ , _)

L'unfolding di **Videocorsi** produce l'OR di 3 query

GAV QUERY UNFOLDING – 2

1. **BooksByVideotutorialAuthor** (author, title, descript, isbn, img, publication_date, page_number) :-
   IsbnDirectory (title, descript, author, isbn, img, publication_date, page_number) ^
   Udemy (_, _ , author, _ , _) ~~^ description = ‘’~~

2. **BooksByVideotutorialAuthor** (author, title, descript, isbn, img, publication_date, page_number) :-
   IsbnDirectory (title, descript, author, isbn, img, publication_date, page_number) ^
   Lynda (_ , _ , author, _ , _) ~~^ price = ‘Free’~~

3. **BooksByVideotutorialAuthor** (author, title, descript, isbn, img, publication_date, page_number) :-
   IsbnDirectory (title, descript, author, isbn, img, publication_date, page_number) ^
   MicrosoftAcademy (_ , _ , _ , _ , _) ~~^ price = ‘Free’~~

LAV BUCKET ALGORITHM – 1

- **BooksByVideotutorialAuthor** (author, title, description, isbn, img, publication_date, page_number) :-
  Books (title, description, author, isbn, img, publication_date, page_number) ^
  Videocorsi (_, _ , author, _ , _ , _)

- bucket [**Books** (title, description, author, isbn, img, publication_date, page_number)]:
  - **IsbnDirectory** (title, description, author, isbn, img, publication_date, page_number)

- bucket [**Videocorsi** (_, _ , author, _ , _ , _)]:
  - **Udemy**(_ , _ , author, _ , _)
  - **Lynda** (_ , _ , author, _ , _)
  - **MicrosoftAcademy** (_ , _ , author, _ , _)

LAV BUCKET ALGORITHM – 2

Sono possibili 3 riformulazioni:

q1 (author, title, description, isbn, img, publication_date, page_number) :-

IsbnDirectory (title, description, author, isbn, img, publication_date, page_number) ^
Udemy(_ , _ , author, _ , _)

Unfolding:

q1' (author, title, description, isbn, img, publication_date, page_number) :-

Books (title, description, author, isbn, img, publication_date, page_number) ^

Videocorsi (_, _ , author, _ , _ , _)

$$q1' \subseteq BooksByVideotutorialAuthor$$

LAV BUCKET ALGORITHM – 3

Sono possibili 3 riformulazioni:

q2 (author, title, description, isbn, img, publication_date, page_number) :-

IsbnDirectory (title, description, author, isbn, img, publication_date, page_number) ^
Lynda (_ , _ , author, _ , _)

Unfolding:

q2' (author, title, description, isbn, img, publication_date, page_number) :-

Books (title, description, author, isbn, img, publication_date, page_number) ^

Videocorsi (_, _ , author, _ , _ , _)

$$q2' \subseteq BooksByVideotutorialAuthor$$

LAV BUCKET ALGORITHM – 4

Sono possibili 3 riformulazioni:

q3 (author, title, description, isbn, img, publication_date, page_number) :-

IsbnDirectory (title, description, author, isbn, img, publication_date, page_number) ^
MicrosoftAcademy (_ , _ , author, _ , _)

Unfolding:

q3' (author, title, description, isbn, img, publication_date, page_number) :-

Books (title, description, author, isbn, img, publication_date, page_number) ^

Videocorsi (_, _ , author, _ , _ , _)

q3' ⊆ BooksByVideotutorialAuthor

LAV BUCKET ALGORITHM – 5

- La riformulazione LAV di BooksByVideotutorialAuthor è quindi:
  - **q1’** or **q2’** or **q3’**

TECNOLOGIE UTILIZZATE

- PHP
- Tidy
- JavaScript
- MySql
- JSON
- AngularJS
- Angular Material

POSSIBILI SVILUPPI FUTURI

- Aggiornamento automatizzato della cache con tutte le ricerche nella lista dei suggerimenti. -> Ricerche più veloci
- Integrare ulteriori fonti per la ricerca sui prezzi dei libri, degli eventi e delle communities.
- Estrarre dalle pagine del dipartimento i testi utilizzati nei vari corsi e cercare automaticamente i prezzi.
- Integrare Apache Solr per il full-text indexing dei tutorial per la ricerca tramite keyword