Università della Calabria

Dipartimento di Matematica e Informatica
Corso di laurea in Informatica

Data Warehouse Project
CORDIS - EU research 2007-2013

Giovanni Brunetti
*Matricola:*193452

Anno accademico 2017/2018

Introduzione

Lo scopo del progetto è quello di costruire un sistema di Data Warehouse basato sui dati relativi all'assegnazione di progetti scientifici da parte dell'UE alle varie organizzazioni mondiali.
Per far ciò si provvederà alla creazione del/dei data-mart necessari per i fatti di interesse in modo da permettere lo svolgersi delle analisi OLAP su di essi. La progettazione ha seguito le varie fasi relative all'approccio **source oriented**, includendo lo studio delle sorgenti, comprendendo operazioni di pulizia e di trasformazione dei dati per comprendere i dati ed il dominio di interesse. In seguito si è passato alla fase di alimentazione del database e dello studio dei dati ottenuti tramite analisi.

Per la fase di studio di ETL ed alimentazione del DB si è utilizzato il software Pentaho, mentre per lo studio dei dati e l'effettuazione delle analisi si è utilizzato il software Tableau.

Sorgente dei dati

La sorgente dei dati è reperibile all' url
`https://data.europa.eu/euodp/data/dataset/cordisfp7projects` sotto forma di file csv.
La sorgente è composta da 6 file, essi contengono tutte le informazioni relative ad ogni singolo progetto assegnato dal 2013 al 2017. La prima fase del progetto si è dunque concentrata nello studio di tutti i file per capire come essi erano connessi tra di loro e come esprimevano il dominio di interesse. Essi presentavano molte ripetizioni di attributi nei diversi file con conseguente rindondanza di informazioni. Dopo uno studio e la selezione degli attributi più rilevanti per il data mart scelto informazioni espresse e rilevanti sono:

- nome del progetto;
- le organizzazioni coinvolte nel progetto;
- la nazionalità di ogni organizzaione;
- programma di appartenenza del progettto;
- contributo economico dell'ue per ogni organizzazione;
- data di inizio e di fine.

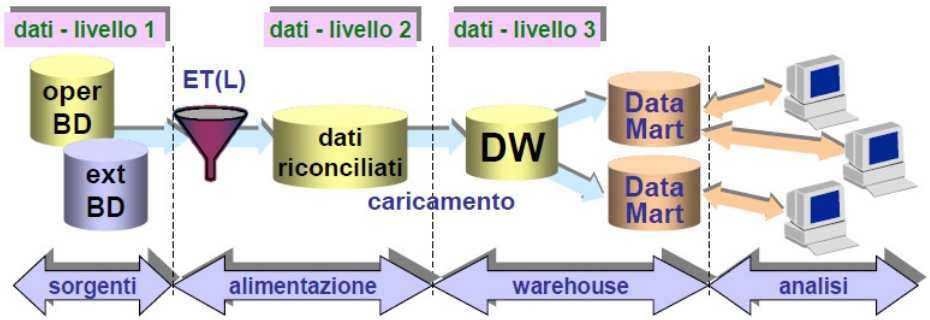
Architettura di sistema

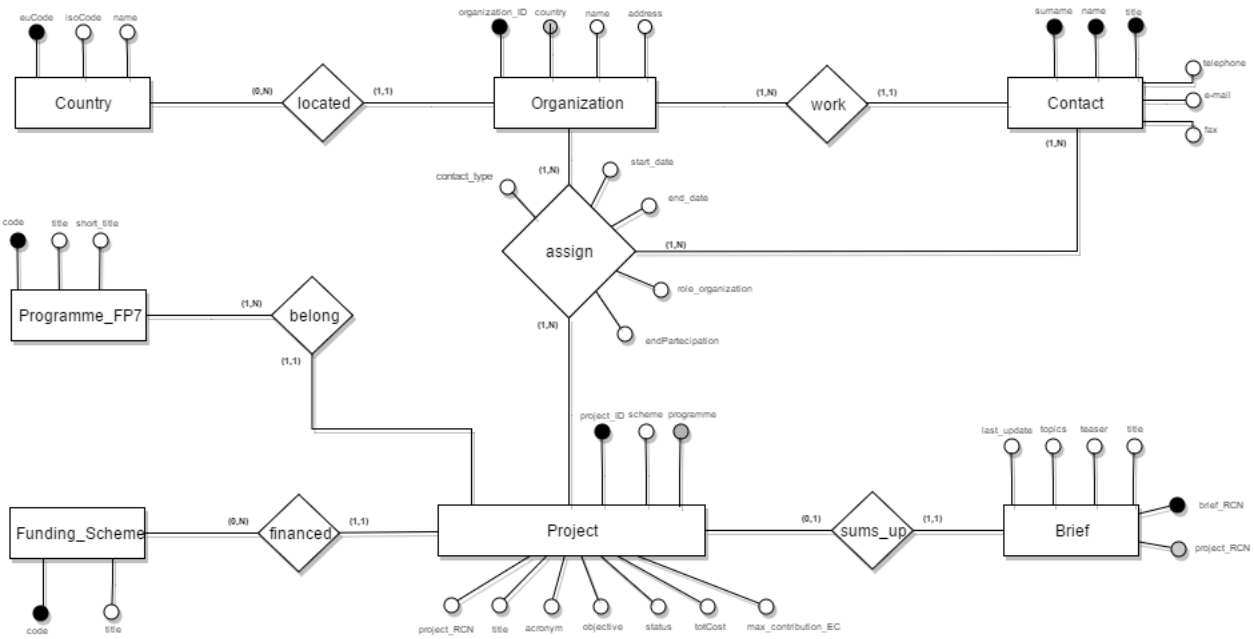
L'architettura di sistema scelta per l'implementazione è quella a 3 livelli. Questa scelta è dovuta al fatto di avere più file come sorgente. Il secondo livello, ovvero quello dei dati riconciliati, permette infatti di integrare le molteplici sorgenti per definire un'unica fonte di dati operazionali. Così facendo si divide la fase di pulitura e di **ETL** nel DW, con la conseguente creazione delle dipendenze funzionali, da quello dell'alimentazione del Data Warehouse.

Database riconciliato

Per la creazione del livello riconciliato è stato utilizzato un modello E/R. Prima di ciò è stato effetuta la fase di **Ricognizione** sul dominio applicativo per costruire le dipendenze funzionali tra i dati iniziali. Lo schema E/R ottenuto è il seguente:

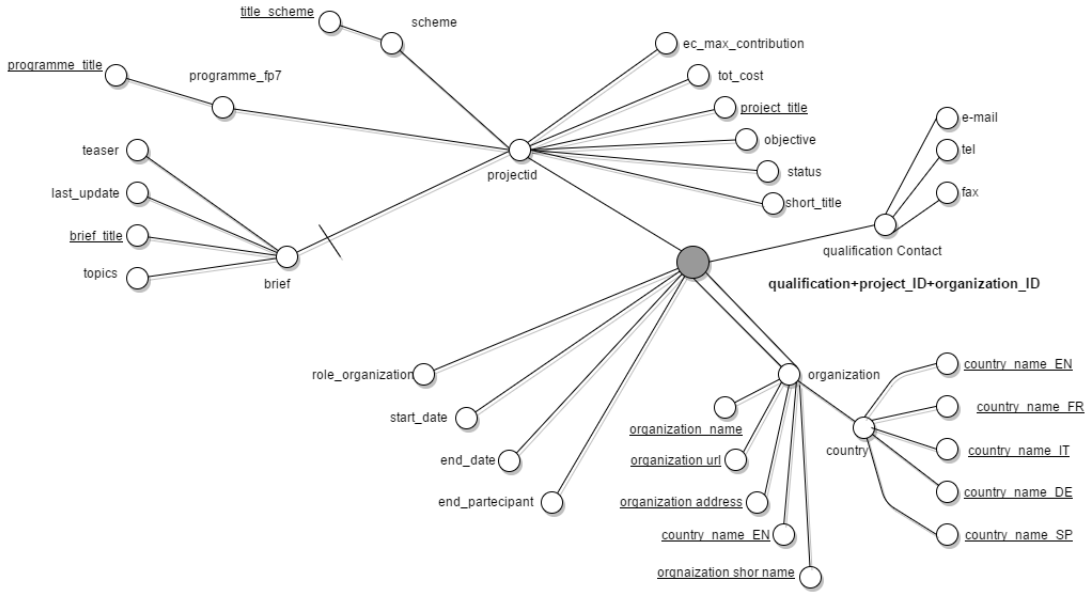
Come si può subito notare il fatto di interesse (**Data Mart**) è quello relativo all'assegnazione dei progetti, **assign**.


dati - livello 1
dati - livello 2
dati - livello 3
oper BD
ext BD
ET(L)
dati riconciliati
DW
caricamento
Data Mart
Data Mart
sorgenti
alimentazione
warehouse
analisi

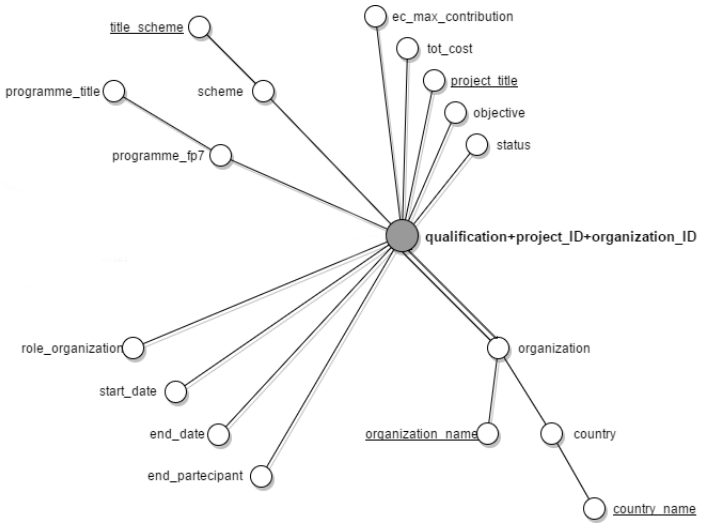

euCode
isoCode
name
Country
(0,N)
located
(1,1)
organization_ID
country
name
address
Organization
(1,N)
work
(1,1)
surname
name
title
Contact
telephone
e-mail
fax
(1,N)
contact_type
(1,N)
start_date
end_date
assign
(1,N)
role_organization
(1,N)
endPartecipation
code
title
short_title
Programme_FP7
(1,N)
belong
(1,1)
project_ID
scheme
programme
Funding_Scheme
(0,N)
financed
(1,1)
Project
(0,1)
sums_up
(1,1)
last_update
topics
teaser
title
Brief
brief_RCN
project_RCN
code
title
project_RCN
title
acronym
objective
status
totCost
max_contribution_EC

Progettazione Concettuale

La prima fase per la progettazione concettuale di un data mart è quello di definire il fatto di interesse, cioè il concetto primario su cui verranno fatte tutte le analisi e le operazioni. Come già detto in precedenza il fatto scelto in questo progetto è l'assegnazione vera e propria del progetto. Si costruisce così l'**albero degli attributi**, attraverso il quale, partendo dal fatto come radice, si delimita l'area di interesse dello schema inserendo come figli tutte le dipendenze funzionali in modo ricorsivo.



In seguito l'albero viene ristrutturato tramite pruning, operazioni sulle dipendendenze funzionali (potatura e innesto) ed eliminando gli attributi irrilevanti ai fini dell'analisi.

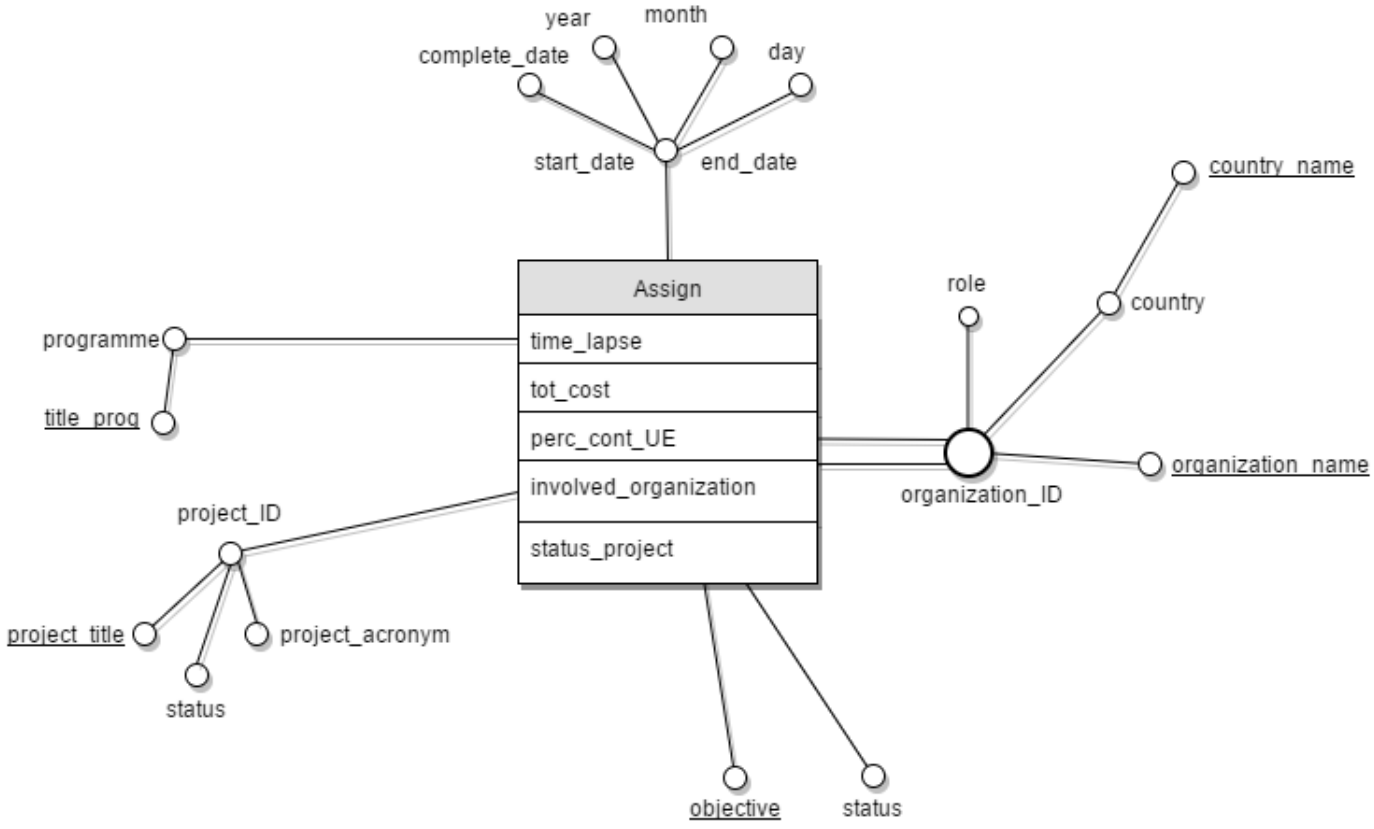


Le operazioni effettuate sono le seguenti:

- pruning del ramo brief e qualification, in quanto irrilevante per le analisi;
- innesto del solo qualificatore qualification in assign;
- potatura attributi insignificanti per le analisi (solo descrittivi o insignificanti);
- innesto attributi del sotto albero project direttamente al fatto assign.

Una volta ottenuto l'albero degli attributi si passa alla definizioni delle **dimensioni**, le quali permetteranno l'aggregazione degli eventi, e delle **misure**, che di fatto definiscono dei criteri di valutazione del fatto. Le dimensioni scelte sono programme, organization e start_date e end_date.
Con la definizione di misure e dimensioni si può passare alla creazione dello **schema di fatto**, composto dal fatto centrale, con le misure al suo interno, e dalle **gerarchie**, che corrispondono ai rami dell'albero degli attributi aventi come radice una dimensione.

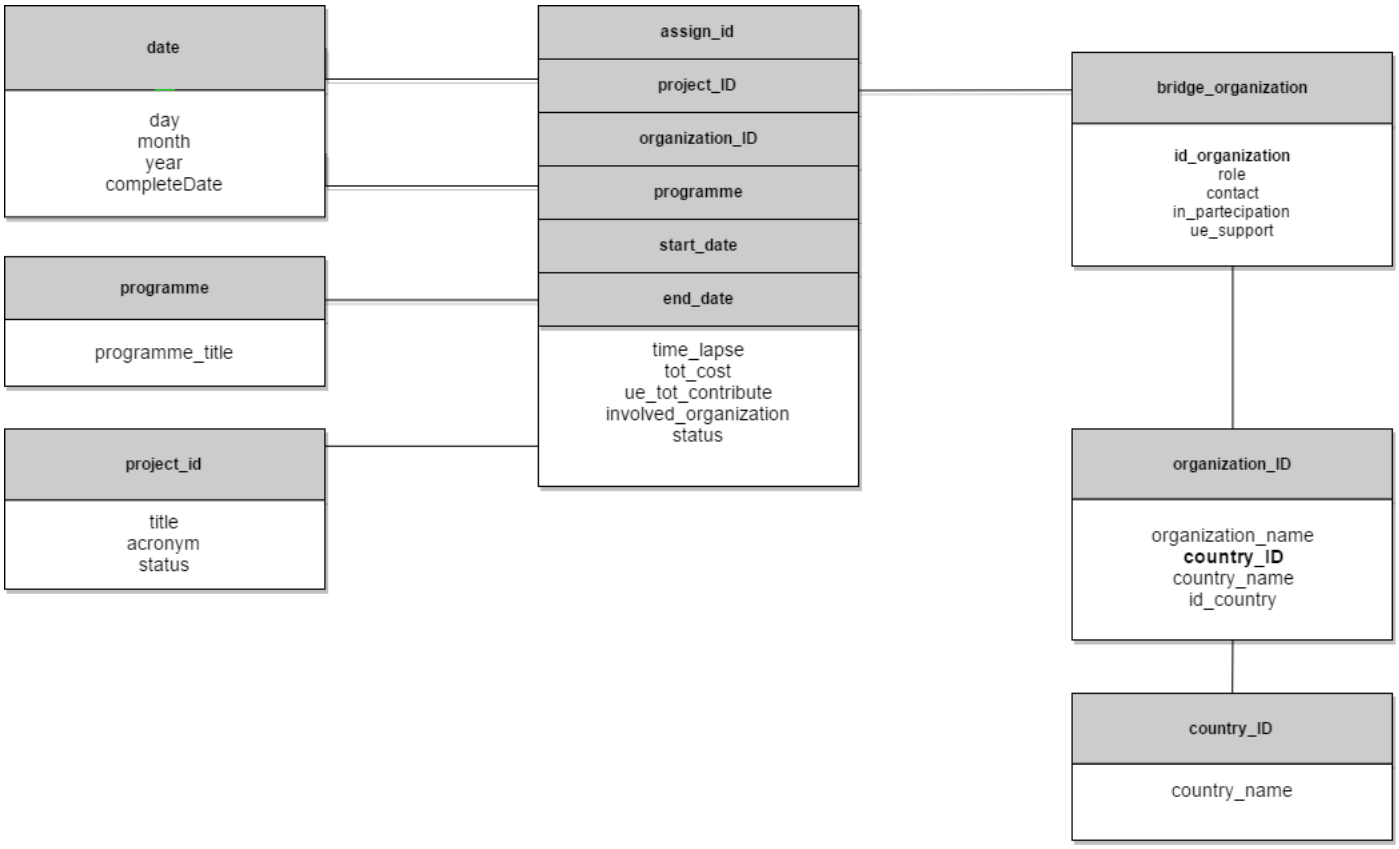
Durante questa fase si è potuto applicare nuove operazioni di modifica, quali potatura e innesto. Si è effettuato l'innesto degli attributi year, month e day alle dimensioni date (start e end). Per comodità è stata usata la stessa radice per start e end, poiché avevano una gerarchia condivisa. Inoltre per facilità di aggregazione, project è stato riinnestato come dimensione, per permettere un'analisi per singolo progetto in modo più semplificato.



MISURE SCELTE CON RELATIVE FUNZIONI DI AGGREGAZIONE:

sum(time_lapse)
sum(status)
sum(perc_ue_cont)
sum(involved_organization)
tot_cost

Una volta creato il fact model non resta che creare lo **star_schema**, che altro non è che la rappresentazione, attraverso tabelle dello stesso fact model. Si costruisce



Si noti che è stato necessario la gestione dell'**arco multiplo** tra la dimensione organization e il fatto. Ciò è stato risolto con l'inserimento di una **bridge_table**.
Si è scelto di utilizzare questo approccio, invece che uno **push_down**, in modo da non duplicare istanze all'interno della fact table e risparmiare dunque memoria. Infatti il metodo push_down consisteva nell'inserire più eventi con le combinazioni dell'attributo multiplo all'interno del fatto stesso. Nonostante sia vero che il metodo push down garantirebbe una maggiore velocità, il metodo della bridge_table garantisce una maggiore facilità nelle analisi d'impatto.

Alimentazione

L'alimentazione del Data Warehouse è avvenuto basandosi sul riconciliato. Infatti si è fatto in modo che le operazioni di modifica e di aggiornamento possano avvenire solo all'interno del riconciliato.

Perciò sono stati utilizzati dei dati temporali per verificare l'eventuale modifica (campo **last_update**) delle tuple. Ciò è bastato in quanto le dimensioni appartengono tutte a **Slowly Changing di Tipo 1**, in cui non vi è bisogno di tenere traccia dei precedenti valori dei campi aggiornati, ma solo l'ultima versione.

L'alimentazione di Data warehouse avviene dunque tramite **ETL-update** ed **ETL-refresh**. Il primo consiste nell'effettuare l'aggiornamento dei dati in modo periodico, ad esempio settimanalmente, prelevando i dati dal database riconciliato.
Il secondo invece consiste nel caricamento dei dati in modo statico quando il data warehouse è vuoto.

Analisi effettuate

Si sono effettuate le seguenti analisi per capire la partecipazione delle varie organizzazioni e prelevare informazioni dai progetti:

- Dashboard generale relativa ad una singola organizzazione;
- Dashboard generale relativa all'insieme delle organizzazioni di una o più nazioni;
- Varie analisi singole sulle organizzazioni o su gruppi di organizzazioni;
- Dashboard generale relativa ad una singola nazione o un sottoinsieme di essi;
- Varie analisi sulle organizzazioni;
- Visuale su mappa dell'impegno delle singole nazioni;
- Dashboard generale relativa ai progetti, con possibilità di filtro su uno solo o più;
- Varie analisi sui progetti, quali media fondi, tempo previsto ecc;
- Dashboard generale basata sul tempo, dunque su data di assegnazione e di chiusura per verificare il trend temporale;

Sulla maggior parte delle analisi è possibile applicare determinati filtri per restringere le analisi in un lasso di tempo, a determinate nazioni o ad un certo valore. Ad esempio potremmo essere intenzionati ad analizzare tutte le nazioni che hanno preso parte a più di 100 progetti nell'arco temporale dal 2012 al 2013.