UNIVERSITÀ DELLA CALABRIA

Department of mathematics and computer science

BUSINESS INTELLIGENCE AND ANALYTICS

DATA WAREHOUSE PROJECT

A.A. 2017/18

Prof. Giorgio Terracina

Studente: Alessandro Scarlato matr.193471

pentaho

+ableau

Sommario

1.Introduzione 3

2.Architettura logica 4

3.Caricamento risorse su MySQL 5

4.Progettazione concettuale 6

5.Progettazione logica 7

6.Progettazione dell'alimentazione 7

7.Analisi 7

pentaho

+ a b | e a u

1.Introduzione

L'obiettivo di questo lavoro è quello di progettare un sistema di data warehousing che utilizza come fonte di analisi alcuni dataset estratti dal sito TMDb. In particolare ho utilizzato 4 file sorgenti: un file è costituito dai dettagli dei film, l'altro è costituito dai crediti dei film, uno contiene la corrispondenza delle società che producono film, infine l'ultimo contiene tutti i continenti delle rispettive regioni. Dall'analisi di questi file è stato possibile progettare un database relazione dal quale verranno successivamente estratti e caricati i dati che faranno parte del Data Warehouse. Dopo una prima fase di studio delle sorgenti ed una prima pulizia dei dati "sporchi", ho ricavo il modello E-R e successivamente l'albero degli attributi, come radice è stato posto il concetto di interesse per il processo decisionale ossia "film". Dopo aver apporato alcune modifiche, potature e innesti all'albero è stato dedotto il relativo DFM, detto anche schema dimensionale e consiste di un insieme di di schemi di fatto i cui elementi sono: le misure che descrivono quantitativamente il fatto, le dimensioni che determinano la granularità minima adottata per rappresentare un fatto e le gerarchie che descrivono come le istanze di un fatto possono essere aggregate e selezionate significativamente per il processo decisionale. Infine sono passato alla modellazione logica ed alla realizzazione dello Star Schema.
Tramite l'uso di Pentaho/Kettele sono stati estratti i dati sorgenti, trasformati ed infine caricati nel Data Warehouse, anche l'alimentazione del Data Warehouse è stata gestita tramite Penthao, ed in particolare è stato fatto uso delle slowly changing dimension di tipo 1, in cui si sovrascrive i vecchi dati con i nuovi, e quindi non tiene assolutamente traccia dei dati storici, metodo più appropriato per correggere alcune tipologie di errore sui dati. Le Slowly Changing Dimensions sono dimensioni i cui attributi hanno valori che possono variare lentamente nel tempo. Per concludere sono state eseguite alcune analisi di interesse tramite il software Tableau.

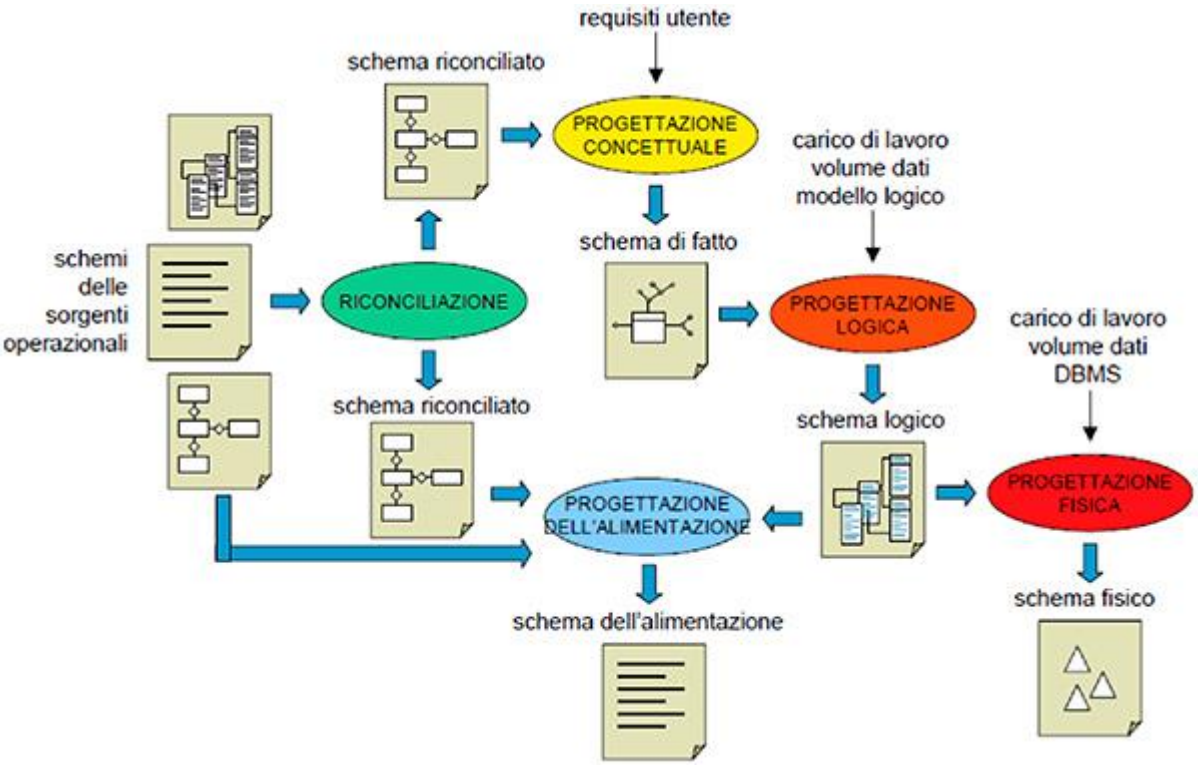

Figura 1. Ciclo di sviluppo di un data warehouse.

2.Architettura logica

Si è scelto di utilizzare l'architettura a tre livelli, in questo caso il DW viene alimentato non più direttamente dalle sorgenti, ma dai dati riconciliati. Essa in realtà si articola su quattro livelli distinti, che descrivono stadi successivi del flusso dati:
1) Livello delle sorgenti: prevede l'estrazione dei dati.
2) Livello di alimentazione: contiene una serie di processi che servono per trasformare i dati, noti con il nome di strumenti ETL (Extract, Transform, Load), il cui ruolo è quello di alimentare una sorgente dati singola, dettagliata, esauriente e di alta qualità che possa a sua volta alimentare il DW (riconciliazione). Durante il processo di alimentazione del DW, la riconciliazione avviene in due occasioni: quando il DW viene popolato per la prima volta, e periodicamente quando il DW viene aggiornato.
3) Livello del warehouse. Le informazioni vengono raccolte in un singolo "contenitore" centralizzato logicamente: il DW.
4) Livello di analisi. Permette la consultazione efficiente e flessibile dei dati integrati a fini di stesura di report, di analisi, di simulazione.

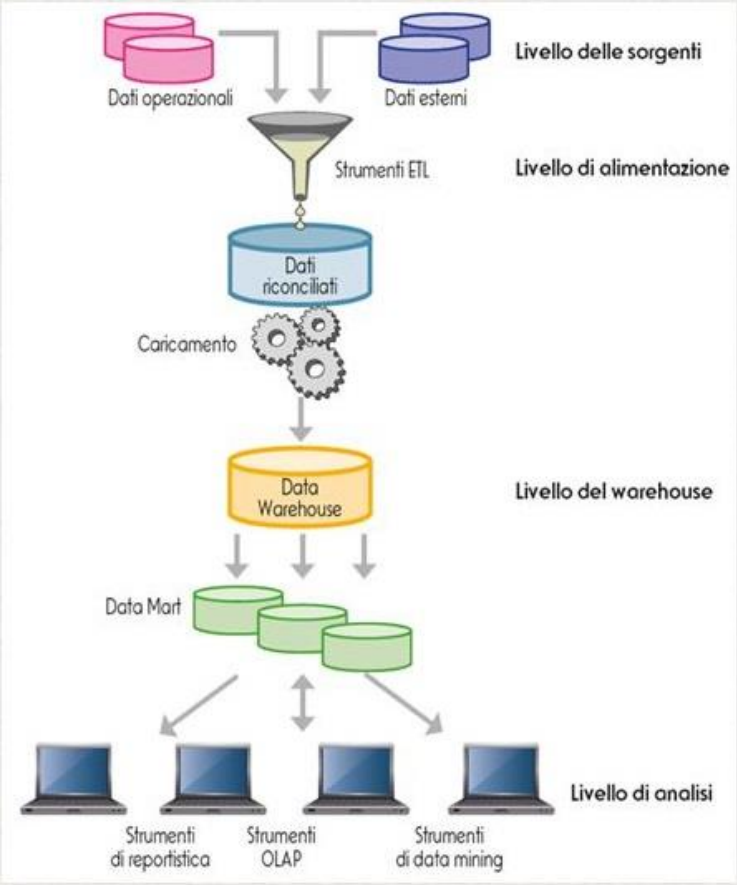

Figura 2. Architettura a tre livelli.

3.Caricamento risorse su MySQL

Le risorse inizialmente in formato csv, sono state divise in entità e caricate sul database (passo essenziale per implementare etl-update).

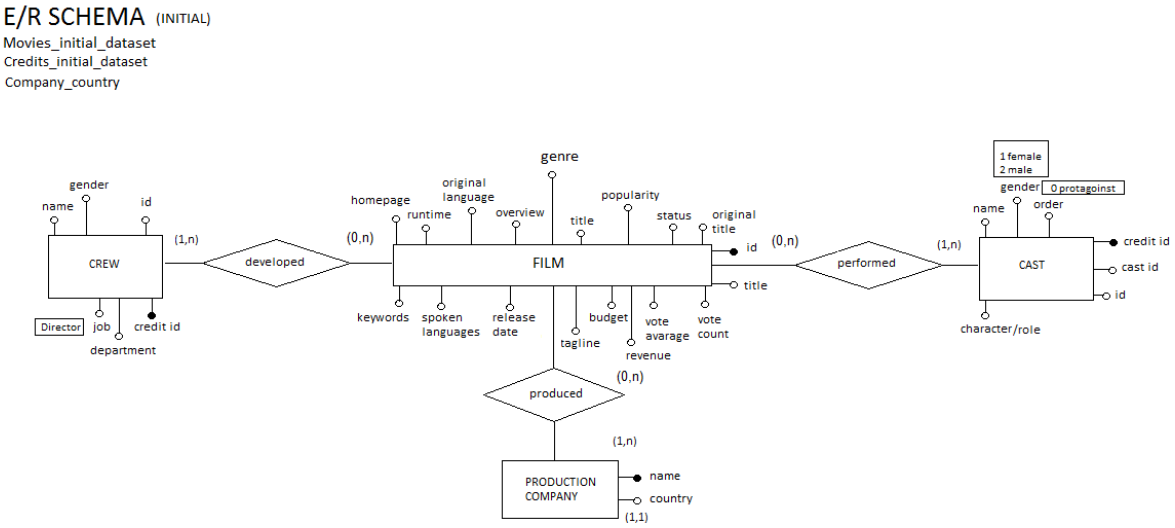

Figura 3. Schema E-R iniziale

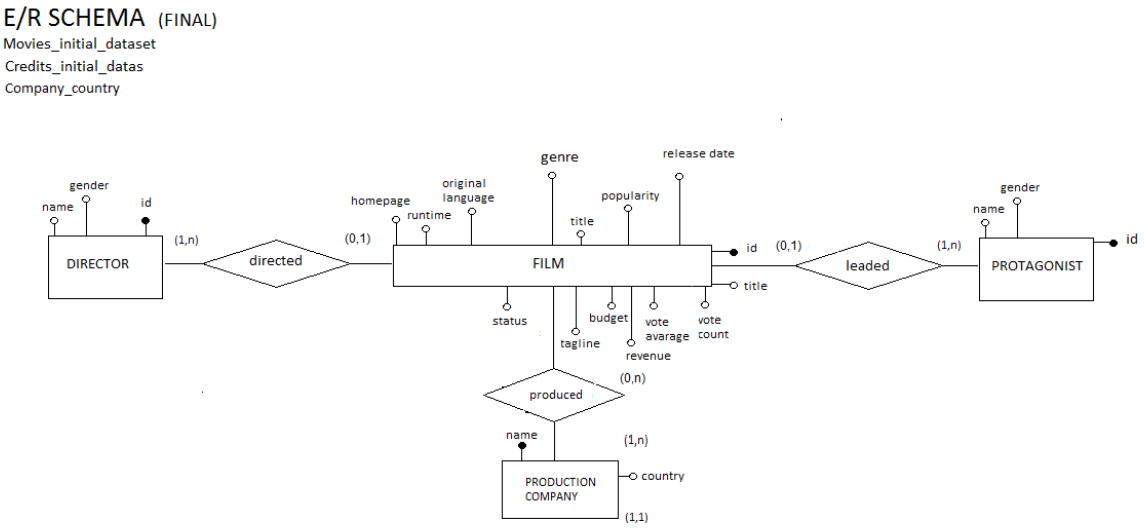

Figura 4. Schema E-R finale

4.Progettazione concettuale

Partendo dallo schema E-R, sono stati realizzati l'albero degli attributi e successivamente il DFM, identificando il fatto le dimensioni e le misure.

ATTRIBUTE TREE INITIAL

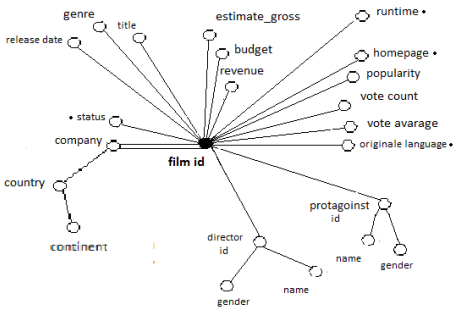

Figura 5. Albero degli attributi.

ATTRIBUTE TREE INITIAL

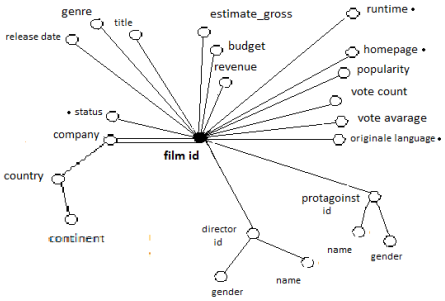

Figura 6. Albero degli attributi finale.

DFM (DIMENSIONAL FACT MODEL)

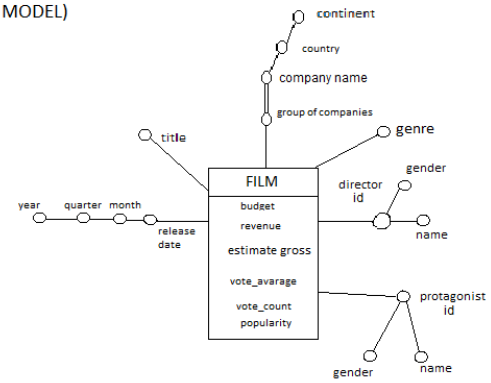

Figura 7. Schema DFM.

5.Progettazione logica

In questa fase si è realizzato lo Star schema

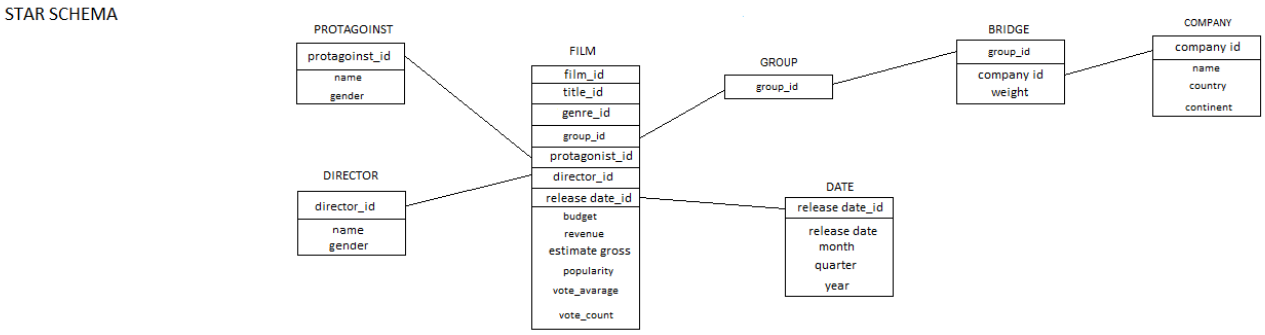

Figura 8. Star schema

6.Progettazione dell'alimentazione

I passi dell'alimentazione sono:
- Estrazione
Il primo passo è stato l'estrazione dei dati dalle sorgenti csv (estrazione statica) nell' area di staging in Pentaho.
- Pulitura e Trasformazione
Successivamente all'estrazione, è stata effettuata la pulitura e tutti i dati sono stati convertiti dal formato sorgente a quello del DW, ottenendo il ivello riconciliato.
-Caricamento
Con tale passo si è ottenuto il livello del Warehouse.
1. ETL-Refresh: Questa tecnica normalmente è utilizzata per popolare inizialmente il DW, ma può essere utilizzata anche per riscrivere integralmente il DW, sostituendo quelli precedenti, per tale ragione è stata aggiunta anche questa possibilità.
2. ETL-Update: Per l'aggiornamento del DW, è stata utilizzata la tecnica etl-update, ovvero vengono prelevati i soli cambiamenti occorsi nei dati sorgente, cioè le tuple che sono state inserite o aggiornate dopo l'ultimo aggiornamento del DW. Questo è stato possibile aggiungendo un campo "stamp" per ogni tabella sorgente, e una tabella last_update che indica l'ultimo aggiornamento eseguito per il DW, viene inizializzato al momento del refresh e cambia ad ogni aggiornamento.

7.Analisi

Attraverso l'utilizzo di Tableau, sono stati sheet, dashboard e una story:
inizialmente, sono state fatte delle analisi di tipo decisionali, utili per la produzione di un film (sottolineati in blu), cercando il miglior genere da produrre; il mese in cui rilasciarlo; quale casa produttrice, l'attore e il regista da contattare.
Infine, ho fatto ulteriori analisi di tipo puramente statitstico.