INTEGRAZIONE E TEST DI SISTEMI SOFTWARE

A.A. 2018/2019

PROCESSO DI DATA WAREHOUSING

DA DATABASE A BUSINESS INTELLIGENCE

AUTORI:

MAFFEI MARCO, mat. 647101

MOSCATO FRANCESCO, mat. 655691

Abstract

L'intento del caso di studio qui presentato è l'analisi del processo di integrazione dei dati relativi all'anagrafica scolastica nazionale; con la realizzazione di un programma che ha il compito di estrarre i dati scelti, trasformarli e, infine, trasferirli nel Data Warehouse; con l'obbiettivo finale di formulare una relazione dettagliata sui processi e sui metodi utilizzati. In più si esegue un'analisi sui dati presi in considerazione, su più punti di vista, utilizzando un preciso tool che permette la produzione di Business Intelligence. Non verranno forniti consigli decisionali ma solo riflessioni sull'efficienza degli strumenti di Data Warehousing.

Keywords: Database, Datawarehousing, Datawarehouse, Business Intelligence, Power BI.

Introduzione

I dati del sistema OLTP a supporto del caso di studio sono stati prelevati da un portale del MIUR (Ministero dell'Istruzione Università e Ricerca), denominato "Portale Unico dei Dati della Scuola" [1]. È lo strumento che il Miur mette a disposizione dei cittadini per dare concreta attuazione al principio della trasparenza, garantendo così un accesso libero alle informazioni e ai dati della scuola senza autenticazione o identificazione, così come previsto dal *comma 136* della *Legge 107 del 2015*, comunemente denominata la "Buona Scuola". Il database in questione conterrà i dati delle scuole statali e il relativo numero di studenti, diversificati per sesso, corrispondenti agli anni scolastici 2015/2016 e 2016/2017; con esclusione delle province autonome di Bolzano, Trento e Aosta.

Per poter trattare i dati presi in caso, è stato progettato uno script ETL con il compito di estrarre i dati dal database originario, pulirli e trasformarli; in modo che siano significativi per il processo di Data Warehousing richiesto. Tali dati verranno processati con lo scopo di creare Business Intelligence che, attraverso un apposito tool OLAP saranno di supporto ai processi decisionali. La Business Intelligence (BI) può essere definita come il processo di

trasformazione dei dati in informazioni e quindi in conoscenza. La conoscenza si ottiene tipicamente in base alle esigenze del cliente, ai processi decisionali dei clienti, alla concorrenza, alle condizioni del settore e alle tendenze economiche, tecnologiche e culturali generali. [2]

Definizione del processo di DataWarehousing con la specifica della procedura ETL realizzata

Per poter definire il processo di Data Warehousing bisogna innanzitutto definire cos'è il Data Warehouse. Secondo W.H. Inmon:

"E' una raccolta di dati integrata, orientata al soggetto, variabile nel tempo e non volatile di supporto ai processi decisionali" [3]

Nella prima fase, per definizione del processo di datawarehousing, è stata studiata la natura del database e il suo dominio applicativo; quindi è stata anche accertata la qualità dei dati per un adeguata e concreta analisi di BI.

Vista la conformazione del database origine si è optato per la scelta della tipologia MOLAP. La tipologia MOLAP è la più utilizzata e ci si riferisce ad essa comunemente con il termine OLAP [4]. Utilizza un database di riepilogo che ha uno specifico motore per l'analisi multidimensionale e crea le "dimensioni" con un misto di dettaglio ed aggregazioni. Risulta la scelta migliore per quantità di dati ridotte, perché è più veloce nel calcolare le aggregazioni e restituire risultati, ma crea enormi quantità di dati intermedi. Per quanto riguarda gli aspetti negativi vi è l'occupazione elevata dello spazio, con conseguenti cali prestazionali, e l'indisposizione, da parte della metodologia MOLAP, nei confronti degli ambienti dinamici dal punto di vista delle dimensioni. [5]

Il processo ETL potrà essere avviato solamente a seguito di un'autenticazione dell'utente, al fine di controllare gli accessi al Data Warehouse. Dopo aver estratto i dati d'interesse dal database, inserendoli in un file temporaneo ".csv", essi entrano in una fase di limbo, detta 'Area di Staging'. Qui avviene la trasformazione dei dati; vengono effettuate eventuali correzioni ortografiche, avviene la gestione dei campi con elementi non conformi e lo scarto dei campi considerati non necessari ai fini di analisi.

Infine, nell'ultima fase, definita come 'Loading', i dati puliti e concreti verranno caricati nel DW, pronti per essere analizzati per generare la BI.

Queste tre fasi (Estrazione, Trasformazione e Caricamento) costituiscono il processo ETL. [6]

Il processo ETL è stato realizzato attraverso la progettazione di uno script, implementato con l'integrazione di diversi linguaggi di programmazione quali: PHP, SQL, CSS e JavaScript. Sfruttando Oracle MySQL, un Relational Database Management System, per l'immagazzinamento del Data Warehouse.

I dati sono ottenuti in base al fatto d'interesse, per "**fatto**" si intende quei valori numerici che esprimono le variabili o le misure obiettivo dell'analisi del nostro sistema di Business Intelligence. In questo caso di studio, il "fatto" d'interesse è la differenza numerica tra i ragazzi che frequentano ancora la scuola e chi non studia più a livello scolastico. Verranno presi in considerazione solo i ragazzi nella fascia di età 5–19, quindi nell'età scolastica, organizzati per regione.

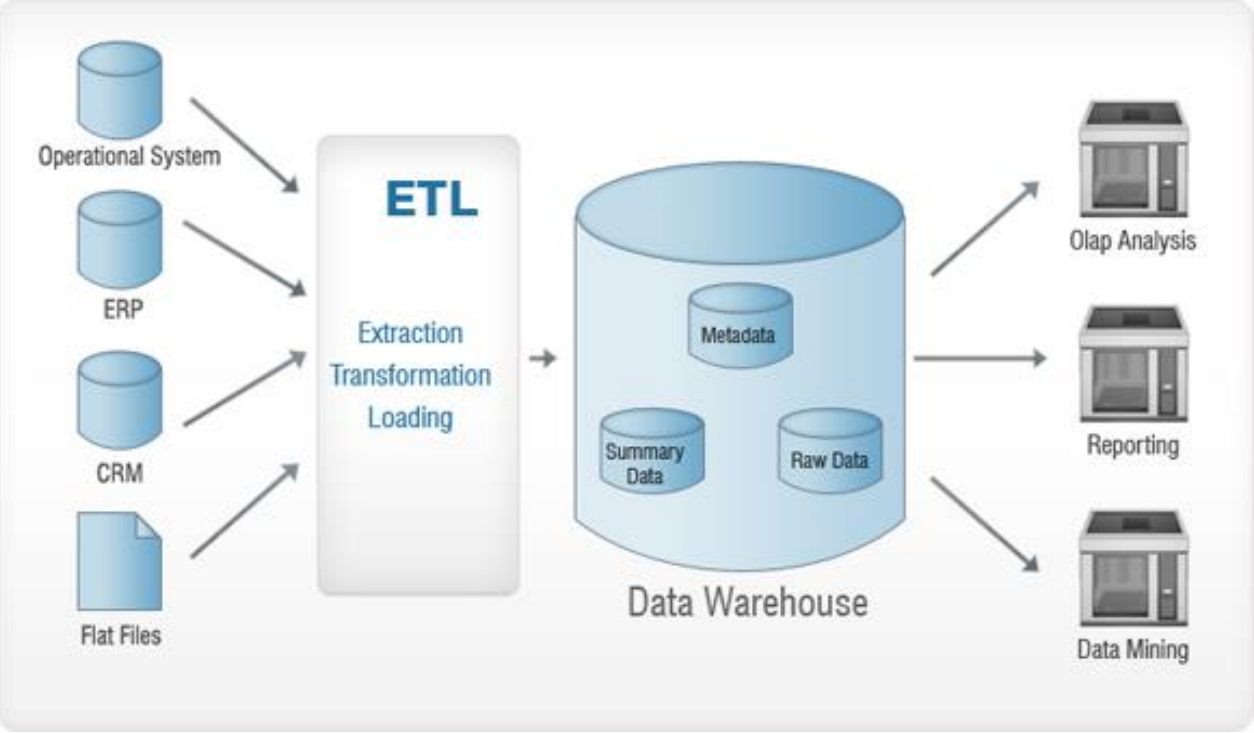

Figura 1: Esempio di un processo di Data Warehousing; (http://www.bidw.ca/tag/datawarehouse-architecture)

Tool utilizzato per le analisi OLAP – Power BI

Power BI rappresenta la famiglia di strumenti di casa Microsoft per la Business Intelligence, grazie ai quali permette di trasformare i dati aziendali in un insieme di informazioni interattive, utili a prendere decisioni. Per questo caso di studio verrà utilizzato "Power BI Desktop", cioè un applicazione desktop su piattaforma Windows per la costruzione dei report BI. La famiglia di strumenti di Power BI si allarga anche per il mobile (Windows, iOS ed Android) e per i servizi online, ma non verranno trattati in questa sede.

La decisione di utilizzare Power BI comporta vantaggi molteplici e riassumibili come segue:

- Prevede l'accesso ad informazioni in tempo reale e approfondimenti su eventi in corso, così da identificare le tendenze del momento.
- È semplice e intuitivo, consentendo di porre domande e ottenere risposte in modo rapido, tramite creazione di diagrammi e grafici.
- È possibile eseguire una ricerca automatica nei set di dati per reperire rapidamente informazioni nascoste e ottenere dati su correlazioni, punti di modifica e altri fattori rilevanti.
- La soluzione Power BI rappresenta un'innovazione continua, per visualizzare e analizzare i dati in modo sempre più veloce, efficiente ed efficace.
- Infine, c'è tutta la forza di una comunità di partner e sviluppatori, tra le più attive, che attraverso Microsoft AppSource offre contenuti e strumenti addizionali molto utili al nuovo utente.

Per "Power BI Desktop", i costi non esistono: la versione è infatti disponibile gratuitamente sul market e permette di accedere ad una serie di funzioni per cui, salvo necessità particolari, non vi sarà bisogno di integrazioni.

Il modo di usare Power BI può variare in base al ruolo ricoperto all'interno di un progetto o di un team. Chi si occupa di vendite potrebbe sfruttare in misura

maggiore l'app per telefoni di Power BI per monitorare l'avanzamento delle quote di vendita e approfondire nuovi dettagli importanti dal proprio dispositivo.

Esperienza

Definito il fatto d'interesse è stato necessario servirsi di una mole di dati che contenesse le informazioni adeguate a tal fine. Il MIUR mette a disposizione, su un suo portale, una vasta gamma di informazioni nell'ambito scolastico.

I dati interessanti erano in particolare due: "Informazioni anagrafiche scuole statali" e "Studenti per anno di corso, classe e genere. Scuola statale anno scolastico". Entrambi i file sono scaricabili in formato ".csv". Per poter utilizzare i dati è necessario ricostruire il database d'origine sfruttando il processo di reverse engineering sulla base dei ".csv".

Quindi il database originato conterrà due tabelle con le seguenti strutture:

Tabella 1: Struttura Dataset Scuole

Attributo	**Tipo**	**Descrizione attributo**
Id	Numerico	Chiave primaria, identifica univocamente il record
AnnoScolastico	Numerico	Anno scolastico di riferimento anagrafe scuola
AreaGeografica	Testo	Area geografica della Regione
Regione	Testo	Regione della Provincia di appartenenza territoriale
Provincia	Testo	Provincia di appartenenza territoriale del Comune ove è sita la scuola.
CodiceIstitutoRiferimento	Testo	Codice dell'istituto a cui fa riferimento la scuola (plesso)

DenominazioneIstitutoRiferimento	Testo	Denominazione (nome) dell'istituto di riferimento della scuola
CodiceScuola	Testo	Codice della scuola (plesso)
DenominazioneScuola	Testo	Denominazione (nome) della scuola (plesso)
IndirizzoScuola	Testo	Indirizzo di recapito della scuola
CAPScuola	Testo	C.A.P. della scuola (plesso)
CodiceComuneScuola	Testo	Codice catastale della scuola (plesso)
DescrizioneComune	Testo	Descrizione codice catastale scuola (plesso)
DescrizioneCaratteristicaScuola	Testo	Descrizione della caratteristica della scuola. Assume valori: ANN. A EDUCANDATO, ANN. C\O IST. OSPEDALIERO, CONVITTO ANNESSO, EDUCANDATO FEMMINILE, ISOLANO, PER CIECHI, PER SORDOMUTI, SPEC. PER CARCERARI, SPEC. PER SORDOMUTI, SPECIALE PER CIECHI, Non Disponibile (se la scuola non ha caratteristiche)
DescrizioneTipologiaGradoIstruzioneScuola	Testo	Descrizione della tipologia o del grado di istruzione della scuola (Primaria, Liceo Scientifico, Istituto d'arte, Convitto…)

IndicazioneSedeDirettivo	Testo	Indica se il codice fa riferimento ad una scuola o istituto sede di segreteria o dirigente scolastico (SI-NO)
IndicazioneSedeOmnicomprensivo	Testo	Se valorizzato con 'Non disponibile' indica che la scuola non fa parte di un Omnicomprensivo. Se valorizzato contiene il codice scuole dell'istituto che è a capo dell'Omnicomprensivo. Un Omnicomprensivo è un'istituzione scolastica che comprende scuole o istituti di gradi di istruzione diversi.
IndirizzoEmailScuola:	Testo	Indirizzo di posta elettronica della scuola
IndirizzoPecScuola:	Testo	Indirizzo di posta elettronica certificata della scuola
SitoWebScuola	Testo	Indirizzo del sito web della scuola

Tabella 2: Struttura Dataset Alunni

Attributi	**Tipo**	**Descrizione Attributo**
AnnoScolastico	Numerico	Anno scolastico di riferimento anagrafe scuola
CodiceScuola	Numerico	Codice della scuola (plesso)
OrdineScuola	Testo	Indica l'ordine scuola (grado di istruzione) della scuola. Non sono rilevati dati relativi alla scuola dell'infanzia.
AnnoCorsoClasse	Numerico	Indica l'anno di corso in riferimento all'ordine scuola e alla classe. Per scuola Primaria da 1 a 5 o 7 nel caso di pluriclasse, ovvero classi dove vengono raggruppati diversi anni di corso. Per

		scuola secondaria di primo grado da 1 a 3. Per scuola secondaria di secondo grado da 1 a 6
Classi	Numerico	Numero di classi
AlunniMaschi	Numerico	Numero di alunni maschi
AlunniFemmine	Numerico	Numero di alunni femmine
TotAlunni	Numerico	Totale alunni per classe

In più, per fini di analisi, è stata prelevata, dal portale ISTAT [7], un'anagrafica dei ragazzi residenti, nella fascia tra i 5 e i 19 anni, di ogni regione per gli anni interessati, cioè il 2015/16 e il 2016/17. Anche in questo caso si è ricorso al reverse engineering, visto che il file era scaricabile solo in formato ".csv".

Tabella 3: Struttura Dataset Regioni

Attributo	**Tipo**	**Descrizione Attributo**
Anno	Numerico	Anno di riferimento della rilevazione
Regione	Testo	Denominazione regione
TotaleMaschi	Numerico	Numero di residenti di sesso maschile
TotaleFemmine	Numerico	Numero di residenti di sesso femminile
Tot	Numerico	Numero di residenti totale nella regione

Il Database, quindi, avrà la seguente struttura con le seguenti relazioni.

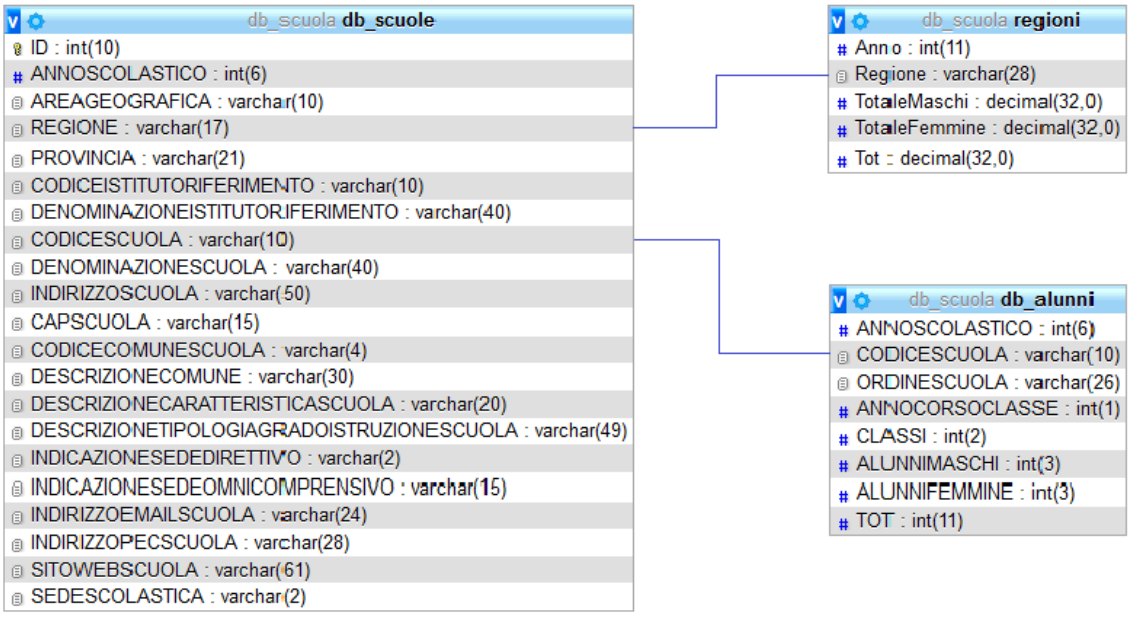

Figura 2: Schema Logico del Database.

Poiché non era disponibile un server online dove far risiedere il database e far girare lo script ETL progettato, è stato utilizzato XAMPP, installato localmente,

per poter usufruire della piattaforma MySQL per la gestione del database e del server Apache per l'esecuzione dello script. Una volta richiamato lo script ed eseguito si otterrà l'eventuale creazione, nel caso in cui non esistesse il DW, e la popolazione/aggiornamento dei dati.

Il DW realizzato avrà la seguente struttura:

Tabella 4: Struttura Dataset Datawarehouse

ATTRIBUTO	**TIPO**
ID	Numerico
ANNOSCOLASTICO	Testo
AREAGEOGRAFICA	Testo
REGIONE	Testo
PROVINCIA	Testo
CODICESCUOLA	Testo
DENOMINAZIONESCUOLA	Testo
DESCRIZIONECOMUNE	Testo
DESCRIZIONECARATTERISTICASCUOLA	Testo
DESCRIZIONETIPOLOGIAGRADOISTRUZIONESCUOLA	Testo
ALUNNIMASCHI	Numerico
ALUNNIFEMMINE	Numerico
TOTALUNNI	Numerico
MASCHIREGIONE	Numerico
FEMMINEREGIONE	Numerico
TOTREGIONE	Numerico

Per poter ottenere risultati di BI richiesti è stato collegato il DW con il tool Power BI. Attraverso l'esecuzione di query il tool ha generato dei grafici quantitativi rappresentati il fatto.

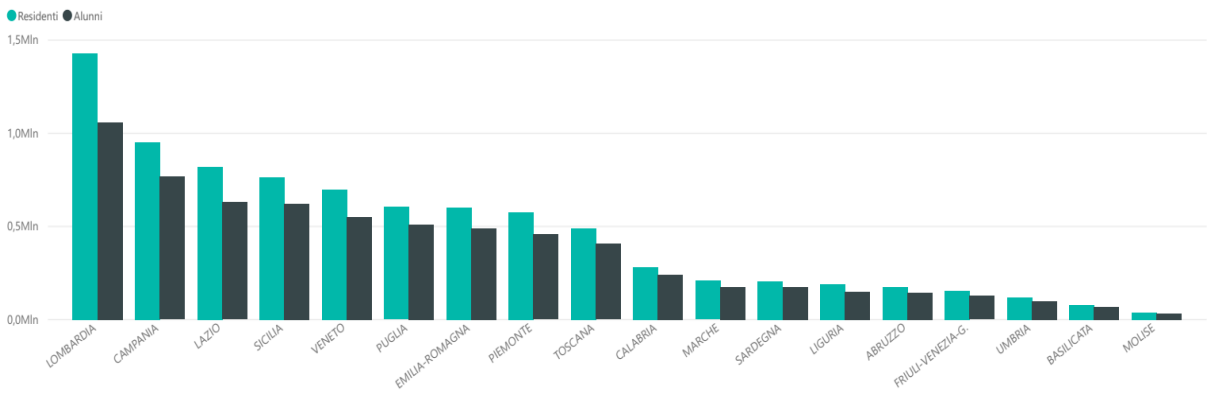

Figura 3: Ragazzi residenti e ragazzi che frequentano qualsiasi tipologia di scuola pubblica

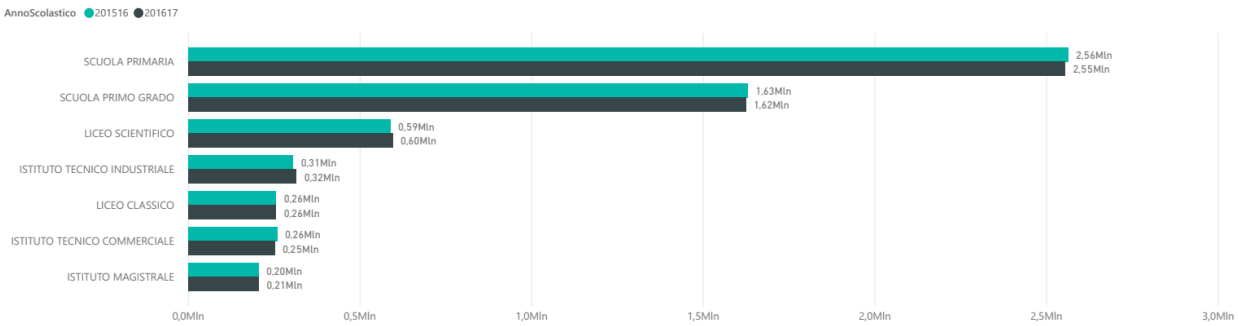

Figura 4: Numero di alunni che frequentano un determinato tipo di scuola categorizzati per anno scolastico

Conclusioni

L'utilizzo del processo di Datawarehousing mette a disposizione dell'utente dati che, se processati attraverso determinati tool OLAP (come Power BI), forniscono grafici quantitativi. La lettura di questi grafici può dare supporto per i processi decisionali anche di elevata complessità.

Un limite del processo di Datawarehousing è la necessità di progettare un processo ad hoc per ogni caso per cui si necessita la produzione di Business Intelligence. Infatti, lo script ETL progettato per questo caso di studio è funzionale solamente per questo dataset.

In base ai risultati ottenuti nel corso di questo caso di studio, si può supporre che le aziende coinvolgeranno sempre di più, nel processo decisionale, la produzione di BI, attraverso i processi di Datawarehousing.

Bibliografia

1. Portale Unico dei Dati della Scuola: http://dati.istruzione.it/opendata/
2. M.Golfarelli, S.Rizzi and I.Cella. "Beyond data warehousing: what's next in business intelligence?" in *DOLAP '04 Proceedings of the 7th ACM international workshop on Data warehousing and OLAP*, 2004.
3. W.H. Inmon. *Building the Data Warehouse*, Wiley, 2005.
4. Y. Zhao, K. Ramasamy, K. Tufte e JF Naughton. *Array-Based Evaluation* e JF Naughton. *Array-Based Evaluation of Multi-Dimensional Queries in Object-Relational Database System*, Proc. 1998 ICDE, pp. 241-249.
5. N. Ferrante. "Datawarehouse delle serie storiche R Italia" in *Bollettino del CILEA n.108*, Roma, 2007

6. C. Koncilia, R. Wrembel. *Data Warehouses and OLAP: Concepts, Architectures and Solutions.* British Library, 2006
7. Demo Istat: http://demo.istat.it/