

Data Mart Segreteria Studenti - Manuale Utente (metadati Pentaho v.01.10.01.00)

- Introduzione
 - Business Intelligence (BI)
 - DatawareHouse (DWH)
 - DataMart (DM)
 - Altre definizioni utili
 - Modello dati multidimensionale (Business Model e Cubo)
 - Metadati
 - Misure o Fatti
 - Dimensioni
 - Gerarchie
 - OLAP
 - Cubo
- Architettura
 - Back-end
 - ETL
 - Staging Area
 - Front-End
- Il Data Mart Segreteria Studenti

Introduzione

In base al principio di autonomia, ogni Università è libera di istituire all'interno del proprio statuto, le regole di governance del proprio Ateneo. Ma che strumenti fornire a tali organi per governare l'Ateneo?

Nella scelta di tali strumenti deve essere preso in esame soprattutto il **processo valutativo** che ha assunto un ruolo centrale per il governo dell'Ateneo negli ultimi anni e che sta prendendo sempre più una forma definita. L'Ateneo deve poter in ogni momento avere a disposizione i principali indicatori che descrivono lo stato di salute dell'Ateneo.

Ed è per questo che, per un Ateneo, assume un'importanza strategica dotarsi di un ambiente di Business Intelligence.

Nel presente volume si vogliono descrivere gli strumenti di Business Intelligence che CINECA mette a disposizione nell'ambito della **Segreteria Studenti**.

Business Intelligence (BI)

Prima di procedere è però necessario fornire alcuni semplici elementi su concetti base quali Business Intelligence, DatawareHouse, DataMart, ODS, rimandando alla bibliografia allegata per ulteriori approfondimenti.

Il termine **Business Intelligence (BI)**, coniato all'inizio degli anni '90 da Howard Dresner, un analista del Gartner Group, viene utilizzato per indicare **un insieme di strumenti e metodologie per la raccolta e l'analisi dei dati, atti a presentare i dati stessi in modo tale che sia possibile estrarne informazioni utili** per gestire i processi decisionali di un'organizzazione.

Si può intuire come questa definizione, piuttosto ampia, includa **varie sottocategorie** tra cui anche il Data Warehousing, Data Mining.

La BI si rivolge principalmente al management (strategico, tattico, operativo) o al knowledge worker, fornendo loro gli strumenti necessari a prendere decisioni e risolvere problemi.

Un sistema di BI non è solo un sistema di reportistica, ma deve avere le seguenti caratteristiche:

- **Facilità d'uso:** presentare i dati in un formato che sia facile da leggere e da interpretare, dove sia possibile navigare sui dati seguendo dei percorsi di analisi e che faccia un ampio uso di grafici. I nomi dei campi devono essere facilmente comprensibili dall'utente finale
- **Velocità:** possibilità di trattare grandi volumi di dati con tempi di risposta quasi istantanei grazie all'uso di tecniche di modellazione, memorizzazione e indicizzazione dei dati orientate all'analisi piuttosto che all'aggiornamento dei dati.
- **Integrazione:** integrare tra loro dati provenienti da fonti differenti, sia interne che esterne all'Ente. Il processo di integrazione deve essere affidabile e testato, in modo che gli utenti possano fare affidamento sui dati presenti nel DW. I dati provenienti dai sistemi gestionali devono passare attraverso un processo di pulizia (data cleansing) e certificazione.
- **Storicizzazione:** mantenere la storia dei cambiamenti subiti da certi attributi selezionati, per permettere analisi storiche contestualizzate.
- **Identificazione di trend ed anomalie:** gli strumenti devono facilitare l'identificazione di trend nei dati, ad esempio confrontando periodi e corsi di studio diversi. Queste operazioni sono possibili solo con l'utilizzo di strumenti interattivi che permettano di effettuare operazioni di drill down/drill up (visualizzazione dei dettagli su un certo dato) e di slice & dice (cambiamento delle dimensioni di analisi sui due assi).
- **Subject orientation:** presentare i dati in modo da fornire la visione di un processo di ateneo attraversando i confini delle singole aree dei sistemi gestionali.
- **Simulazione scenari:** in certi casi deve essere possibile impostare degli scenari e confrontarli poi con i valori reali ("actual")
- **Indipendenza dal reparto IT:** gli strumenti di analisi e reportistica devono dare la possibilità agli utenti finali di crearsi da soli i report di cui hanno bisogno
- **Adattabilità nel tempo,** intesa come la capacità di resistere alle inevitabili evoluzioni della realtà aziendale, dei sistemi operazionali e delle esigenze di analisi
- **Sicurezza:** deve essere possibile controllare in maniera al tempo stesso stretta e flessibile l'accesso ai dati, che in molti casi includono informazioni altamente riservate.
- Soprattutto, nel disegnare un **sistema di BI** è necessario ricordarsi che **deve svolgere bene il suo compito originario, che è quello di supportare le decisioni**.

- Per quanto riguarda la scelta degli strumenti informatici per la realizzazione di un sistema di BI, CINECA ha optato per l'Open Source, utilizzando la **suite Pentaho**.

DatawareHouse (DWH)

Il DataWarehouse è un insieme di dati e strumenti software aventi lo scopo di prelevare i dati dai sistemi gestionali di un'azienda o da fonti esterne e di utilizzarli per effettuare vari tipi di interrogazioni a carattere generalmente statistico/analitico. Per rendere facili e veloci le interrogazioni di grandi volumi di dati, questi ultimi devono prima essere organizzati in maniera differente rispetto ai normali database operazionali (**OLTP** - On Line Transaction Processing).

I DataWarehouse sono generalmente basati sul modello detto dimensionale o **Star Schema**, ottimizzato per rispondere velocemente a interrogazioni di vario tipo.

In generale un DataWarehouse serve ad accrescere la conoscenza di certi fenomeni, che possono riguardare diverse aree di un ateneo (Segreteria Studenti, Stipendi, Contabilità....).

I processi di alimentazione e trasformazione utilizzano dati contenuti nei database operazionali (nel nostro caso le fonti sono ESSE3/GISS e UGOV).

DataMart (DM)

Da quanto sopra esposto, si può facilmente dedurre che, per ottenere informazioni mirate partendo da un sistema di DataWarehouse, è necessario, in alcuni casi, restringere il 'campo visivo' a unità più ridotte e specializzate. Il concetto di DataMart nasce proprio da questa visione e rappresenta un raccoglitore di dati specializzato in un particolare argomento. La progettazione di un Data Warehouse (o Enterprise Data Warehouse), parte solitamente dalla progettazione di singoli DataMart tematici (in ambito universitario, ad esempio, si è partiti dalla progettazione del Data Mart Studenti, Data Mart del Personale ecc...). I singoli DataMart vengono successivamente integrati (processo di federazione dei DataMart) per dare origine al sistema di Enterprise Data Warehouse. Detto in termini più tecnici, un DataMart è un sottoinsieme logico o fisico di un DataWarehouse di maggiori dimensioni.

Altre definizioni utili

Per chi non ha già una conoscenza degli argomenti sopra esposti, in questa sezione definiremo alcuni concetti necessari per la comprensione di quanto sarà esposto nel presente manuale.

Modello dati multidimensionale (Business Model e Cubo)

Organizzazione delle informazioni orientata alle funzioni di reporting e analisi, realizzata per consentire un efficace utilizzo degli strumenti automatici di query, reporting ed analisi. Può essere realizzata fisicamente su strutture dati proprietarie (database multidimensionali) o su database relazionali attraverso una modellazione dati denominata "star-schema".

Metadati

Letteralmente "dati sui dati", è l'insieme di tutte le informazioni che riguardano l'architettura di DWH a parte i dati stessi. Si possono suddividere in **metadati di business**: informazioni a supporto delle attività di analisi (es. data ultimo aggiornamento, anno accademico, corso di studi, unità organizzativa, ecc.)

metadati tecnici: informazioni a supporto dell'attività IT di gestione dell'architettura (es. algoritmi di trasformazione e aggregazione, regole di integrazione, dipendenze dei processi di alimentazione, strutture dati dei sistemi operazionali, responsabili dei sistemi, ecc.).

Misure o Fatti

Valori, generalmente numerici, utilizzati dagli utenti per la misurazione del loro business (es. Numero Iscritti, numero laureati.....).

Dimensioni

Insieme di attributi, generalmente di tipo testo, che definiscono e danno un significato alle misure (es. Tempo, anno accademico, corso di studio.....)

Gerarchie

Naturale correlazione tra attributi appartenenti alla stessa dimensione, dipendente dall'organizzazione e dalle specifiche esigenze applicative.

Possono esistere più gerarchie all'interno della stessa dimensione.

Nei modelli dimensionali le gerarchie stabiliscono le modalità di aggregazione dei dati, ovvero delle misure.

Nel presente datamart NON sono state implementate gerarchie.

OLAP

Letteralmente On Line Analytical Processing, è l'**insieme delle** funzionalità utente che consentono di rendere "dinamica" la visualizzazione delle informazioni attraverso alcune funzionalità utente denominate drill, slice and dice, pivoting :

Drill : Funzionalità OLAP che consente di visualizzare dati a diversi livelli di dettaglio, "navigando" attraverso le gerarchie. Si parla di drill-up quando l'operazione provoca un'aggregazione delle informazioni (es. da "numero iscritti per corso di studi" a "numero iscritti per dipartimento"), drill-down quando succede il contrario.

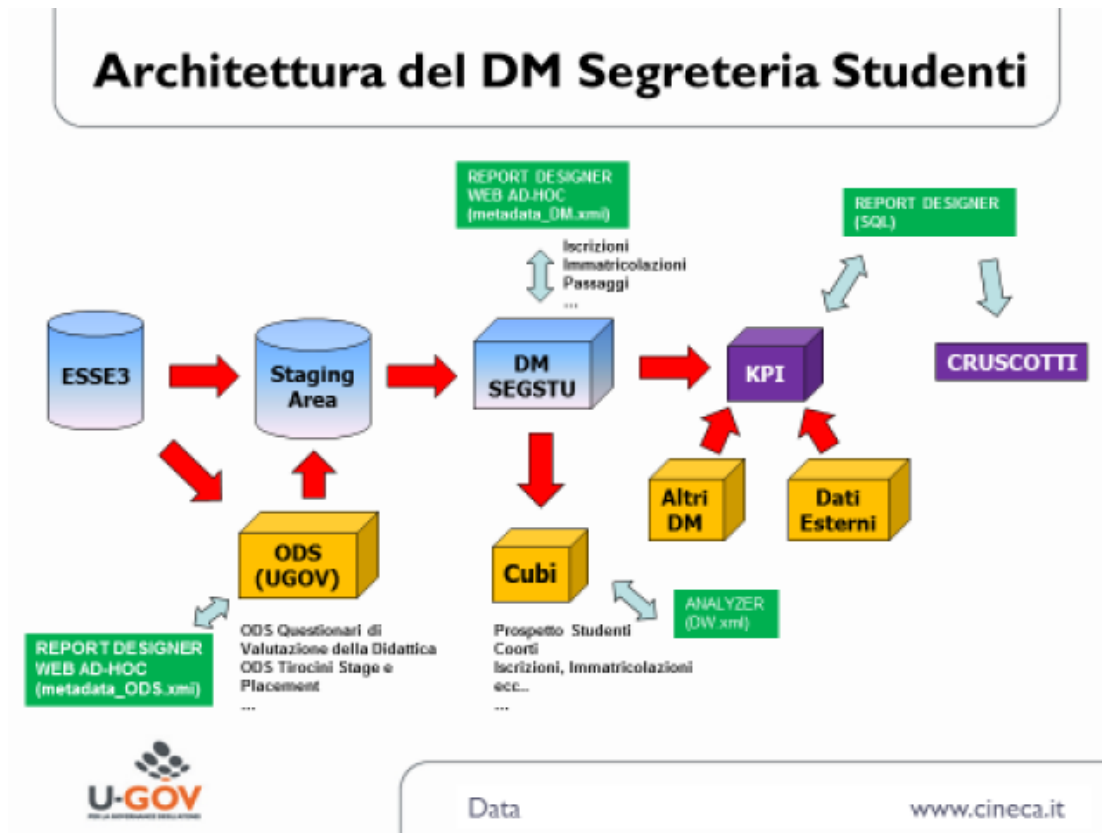
Slice and dice o pivoting : Funzionalità OLAP che consente di ristrutturare le informazioni in modo da renderne più efficace la visualizzazione: creazione di master-detail e rotazione degli assi delle rappresentazioni a matrice.

Cubo

Struttura dati del modello multidimensionale consente una facile navigazione tra i dati su differenti prospettive e diversi livelli di aggregazione. Il CUBO è una struttura dati costruita appositamente per analizzare le informazioni in forma dimensionale.

Architettura

L'architettura del DM SEGSTU è riassunta nella seguente slide. Il presente manuale non tratterà il flusso di alimentazione dei Business Model e dei cubi, ma si concentrerà su di essi, per mettere l'utente in condizione di utilizzare al meglio le informazioni in essi contenute.



I Business Model (BM) a disposizione degli Atenei sono i seguenti:

- STU-Iscrizioni
- STU-Immatricolazioni
- STU-Immatricolazioni Provvisorie
- STU-Abbandoni
- STU-Passaggi
- STU-Trasferimenti
- STU-Esami
- STU-Concorsi
- STU-Lauree
- STU-Piani di Studio
- STU-Crediti

I Cubi a disposizione degli Atenei sono i seguenti:

- STU – ISCRIZIONI, IMMATRICOLATI, IMMATRICOLATI PROVVISORI, ABBANDONI
- STU - ESAMI
- STU - LAUREE
- STU - TASSE
- STU - CONCORSI
- STU - PASSAGGI
- STU - TRASFERIMENTI
- STU - PROSPETTO STUDENTI
- STU - FASCE CREDITI
- STU - COORTE
- STU - PIANI DI STUDIO
- STU – AVA Indicatore 2
- STU – AVA Indicatore 8

Back-end

Con Back-end si intendono i componenti dell'architettura che consentono al DWH di fornire nei tempi e nei modi previsti le informazioni necessarie al business (processi di alimentazione, trasformazione, aggregazione, controllo, ecc.).

ETL

Letteralmente **Extract, Transform, Load**, è l'insieme dei processi che consentono l'alimentazione dell'architettura DWH a partire dai dati contenuti nei sistemi sorgente (operazionali e dati esterni).

Staging Area

E' un'area in cui i dati provenienti dai sistemi operazionali sono temporaneamente memorizzati per essere sottoposti a processi di "pulizia", trasformazione e aggregazione allo scopo di alimentare correttamente il Data Warehouse.

Front-End

Per poter utilizzare, come utenti, i dati di un DataWarehouse o di un DataMart è necessario avere a disposizione un **Front End**, e cioè componenti dell'architettura che consentono la fruizione da parte degli utenti di tali dati. Nel caso del DataMart Segreteria Studenti, il Front End è dato dalla **PUC** (Pentaho User Console) con relativi prodotti specializzati (**Web ad Hoc**, **Web Analyzer** e **Report Designer**)

Il Data Mart Segreteria Studenti

Il DataMart Segreteria Studenti è stato progettato e realizzato per avere dati certificati e controllati nell'ambito, appunto, della Segreteria Studenti.

E' stato progettato e realizzato per la rilevazione di grandezze quantitative e qualitative relative all'efficienza e all'efficacia delle politiche di reclutamento studenti e alle performance degli stessi durante l'intera carriera in Ateneo.

La suite pentaho mette a disposizione due ambienti per modellare i dati: Metadata Editor e Schema Workbench e relativi strumenti per interrogare e elaborare tali dati (nel caso di metadata Editor, Pentaho Report Designer e Web Ad-Hoc; nel caso di Schema Workbench, Analyzer).

I due strumenti sono stati usati per creare due modelli di dati diversi, ma che attingono alla stessa base dati, a seconda delle esigenze espresse dai clienti.

Con metadata Editor sono stati creati i seguenti BM:

- STU-Iscrizioni
- STU-Immatricolazioni
- STU-Immatricolazioni Provvisorie
- STU-Abbandoni
- STU-Passaggi
- STU-Trasferimenti
- STU-Esami
- STU-Concorsi
- STU-Lauree
- STU-Piani di Studio
- STU-Crediti

Come abbiamo già detto, per interrogare tali metadati gli strumenti messi a disposizione degli utenti sono quelli della Pentaho Solution, ovvero

- Pentaho Report Designer
- Web Ad-Hoc

Per le **analisi multidimensionali** sono a disposizione dell'utente i seguenti **CUBI**, creati con **Schema Workbench**:

- STU – ISCRIZIONI, IMMATRICOLATI, IMMATRICOLATI PROVVISORI, ABBANDONI
- STU - ESAMI
- STU - LAUREE
- STU - TASSE
- STU - CONCORSI
- STU - PASSAGGI
- STU - TRASFERIMENTI
- STU - PROSPETTO STUDENTI
- STU - FASCE CREDITI
- STU - COORTE
- STU - PIANI DI STUDIO
- STU – AVA Indicatore 2
- STU – AVA Indicatore 8

In questo caso, lo strumento a disposizione degli utenti è :

- Pentaho Analyzer

