Introduzione ai data warehouse dimensionali

Luca Cabibbo

aprile 2012

Introduzione ai data warehouse dimensionali

Questo corso è un'introduzione ai data warehouse dimensionali, e in particolare alla modellazione dimensionale dei dati – ovvero, alla progettazione logica dei dati – nei data warehouse dimensionali

- contenuti del corso
  - introduzione ai sistemi DW/BI – questo capitolo
  - modello dimensionale e data warehouse dimensionali
  - tecniche di modellazione dimensionale
  - ciclo di vita dimensionale

The Data Warehouse Toolkit

Il corso è basato principalmente sulle tecniche di progettazione di data warehouse dimensionali proposte da Ralph Kimball

- *The Data Warehouse Toolkit (second edition)*
  - Ralph Kimball & Margy Ross
  - John Wiley & Sons, 2002
- *The Data Warehouse Lifecycle Toolkit (second edition)*
  - Ralph Kimball, Margy Ross, Warren Thornthwaite, Joy Mundy & Bob Becker
  - John Wiley & Sons, 2008

Questo materiale è disponibile online al sito

- http://cabibbo.dia.uniroma3.it/dw

Che cosa è un data warehouse

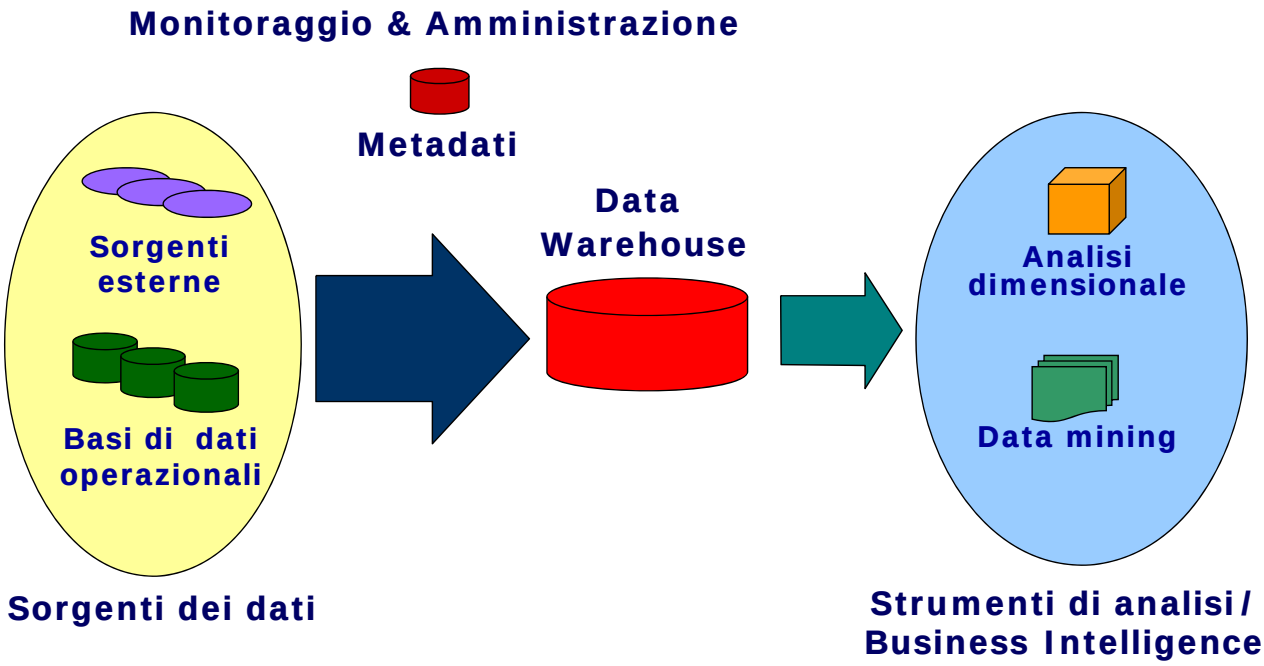
Un **data warehouse** è una base di dati

- orientata ai soggetti
- integrata
- gestita fuori linea
- contenente dati storici
- usata per il supporto alle decisioni direzionali

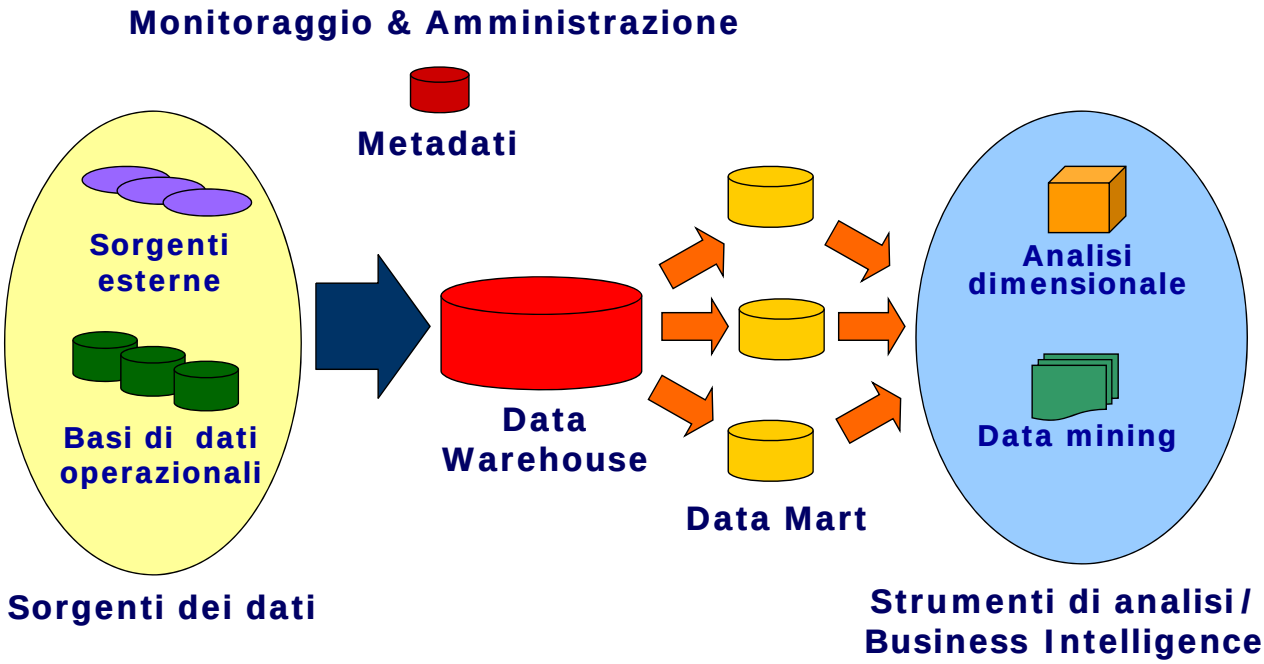
Obiettivi di un data warehouse

- rendere l'informazione aziendale
  - accessibile
  - consistente
  - affidabile
  - sicura
  - usabile per il supporto alle decisioni

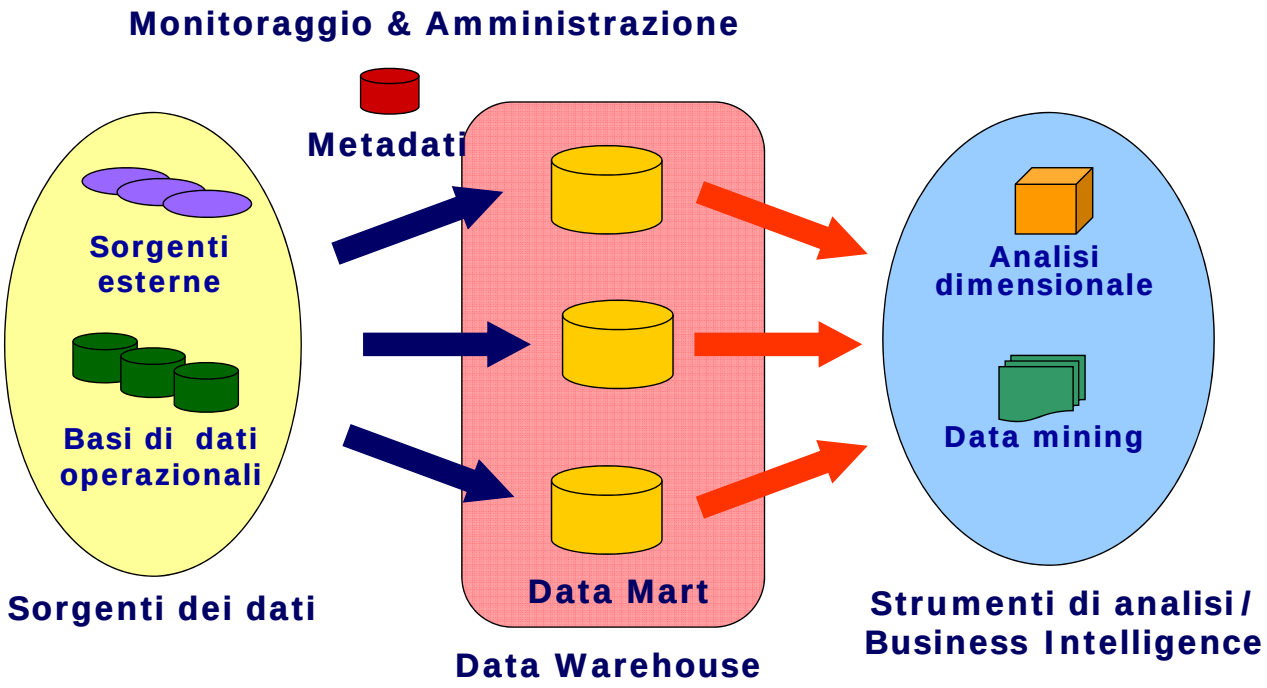
Architettura generale per il data warehousing



Architettura per il data warehousing (Inmon)



Architettura per il data warehousing (Kimball)



Business Intelligence

Business Intelligence (**BI**) è, genericamente, il processo di trasformazione di dati e informazioni in conoscenza

- nello specifico, le tecnologie BI hanno lo scopo di supportare un'organizzazione nello sfruttare al meglio il suo patrimonio informativo (interno ed esterno) nei processi decisionali (decision making)
- "Business Intelligence is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making"
- le tecnologie BI forniscono delle viste storiche, correnti e predittive sui processi di business – alcune funzioni comuni
  - reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining e predictive analytics

Sistemi DW/BI

A questo punto, dovrebbe essere chiaro che un data warehouse ha senso solo come parte di un sistema più ampio

- un **sistema DW/BI** è un sistema completo che racchiude sia il sottosistema data warehouse (DW) che il sottosistema di business intelligence (BI)

Il termine sistema DW/BI enfatizza il legame profondo tra i due sottosistemi – e serve ad evitare possibili incomprensioni

- in teoria, si può fare BI senza un DW
- inoltre, si potrebbe pensare di realizzare un DW senza avere l'intenzione di farci BI – purtroppo questo è successo più volte – ma questo è decisamente sconsigliato

Che cosa è (e non è) un data warehouse

Dunque, una caratterizzazione più precisa di un data warehouse

- il **data warehouse** – **enterprise data warehouse** – è costituito dai dati di un sistema DW/BI
  - in particolare, i dati di un DW possono essere interrogati direttamente dalle applicazioni BI
  - dunque, il data warehouse costituisce le fondamenta per la BI

Viceversa,

- un data warehouse non è semplicemente una base di dati altamente normalizzata, il cui obiettivo principale è servire da sorgente per la trasformazione e il caricamento di dati in strutture dimensionali aggregate – anziché supportare direttamente le interrogazioni delle applicazioni BI

Architettura di un sistema DW/BI

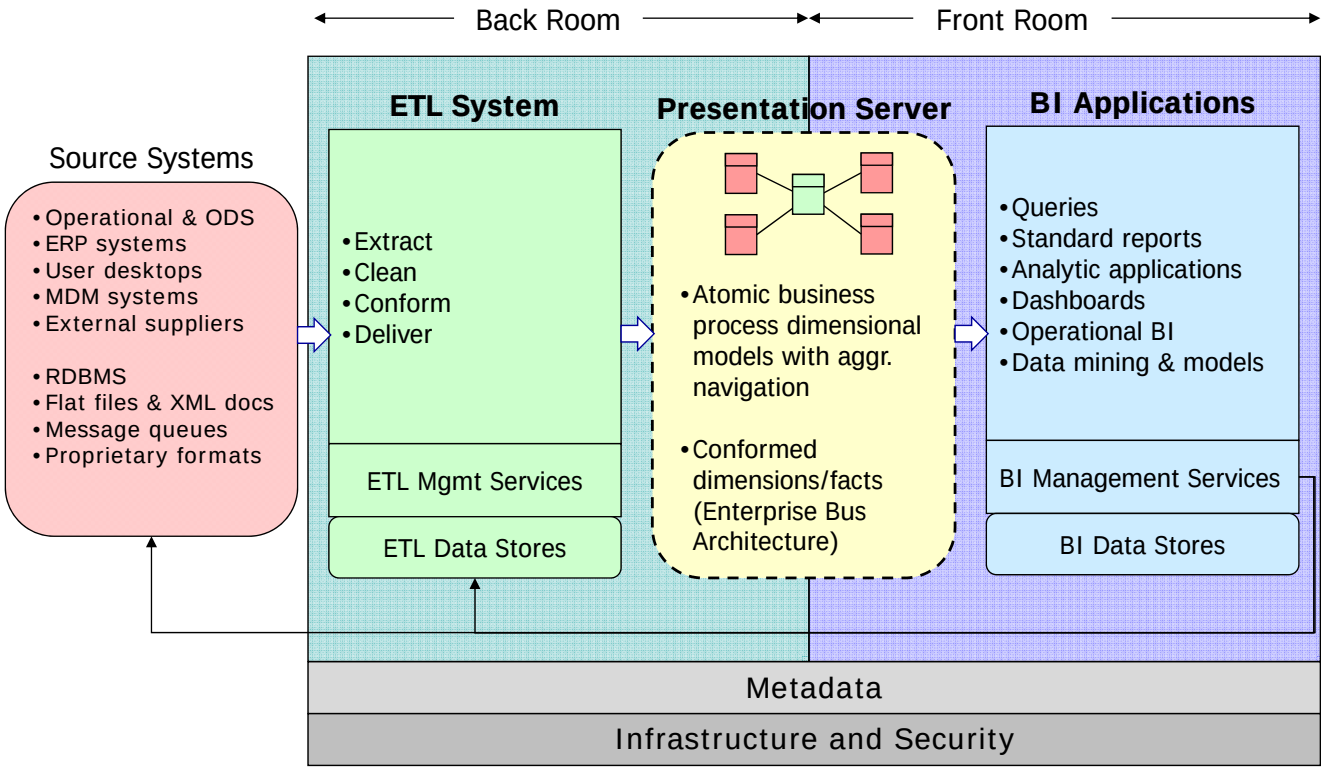
Le funzioni principali di un sistema DW/BI sono

- estrarre i dati dai sistemi sorgenti, pulirli, allinearli, conformarli e trasportali verso il data warehouse
- rendere questi dati disponibili agli utenti di business finali, in modo efficace

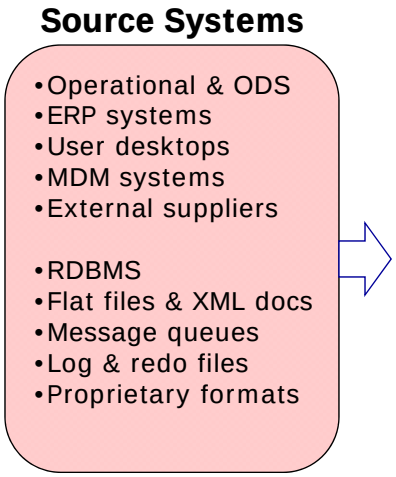
In corrispondenza, l'architettura di un sistema DW/BI conterrà i seguenti elementi principali

- sorgenti di dati (interne ed esterne)
- back room – la parte privata di un DW, dove avviene l'elaborazione ETL
- presentation server – le piattaforme in cui sono memorizzati i dati del DW – questi dati vengono poi interrogati direttamente dal sottosistema BI
- front room – la parte pubblica di un DW – utilizzata direttamente dagli utenti di business del sistema

Elementi di un sistema DW/BI



Sistemi sorgente

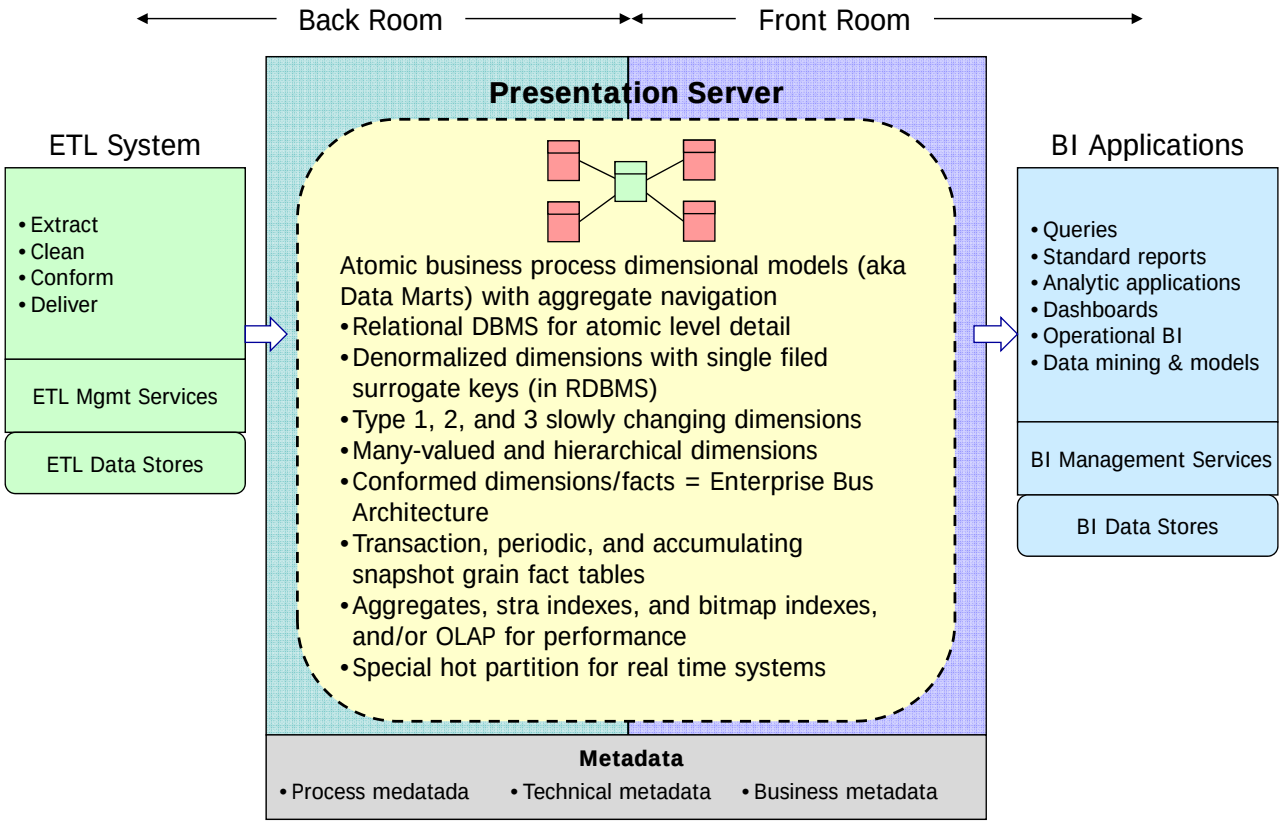


Sistemi sorgente

I **sistemi sorgente** di un sistema DW/BI comprendono di solito numerose *sorgenti informative*

- alcune sorgenti sono i sistemi operazionali dell'organizzazione
  - sistemi transazionali (OLTP) orientati alla gestione dei processi operazionali – gestione ordini, inventario, contabilità, ...
  - di solito non mantengono dati storici
  - possono essere sistemi "legacy"
- altre sorgenti sono esterne all'organizzazione
  - dati pubblici, oppure dati forniti da società specializzate di analisi, spesso relativi al segmento di business di interesse

Presentation server



Presentation server

I **server di presentazione** sono le piattaforme target in cui sono effettivamente memorizzati i dati del data warehouse – per poter essere interrogati direttamente dagli utenti di business finali, dai sistemi di reporting e dalle altre applicazioni BI

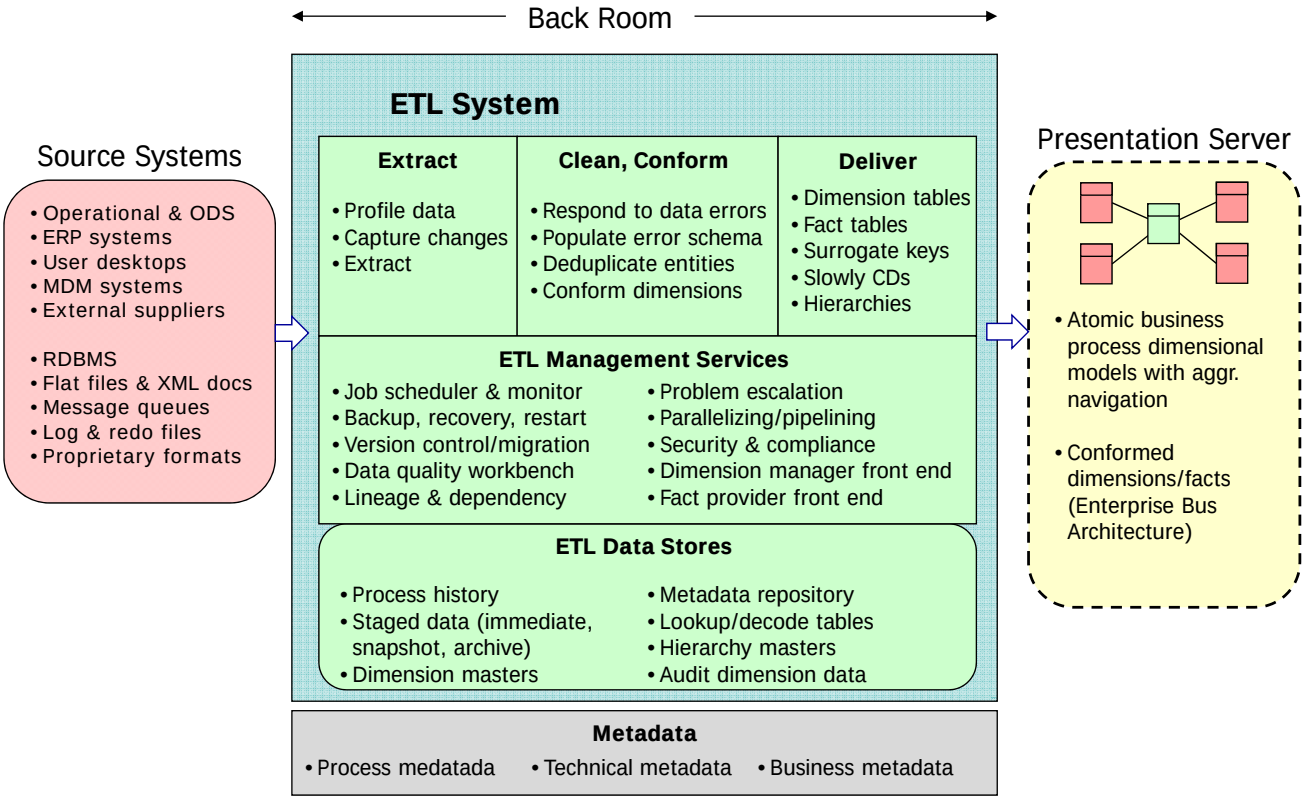
- l'organizzazione dei dati è guidata dalle necessità delle applicazioni BI – che accederanno direttamente a questi dati
  - una singola base di dati per i dati analitici
  - accesso a tutti i dati (di tutti i processi di business) – i dati possono essere visti da punti di vista diversi
  - accesso sia ai dati aggregati che ai dati atomici
- per questo, nei server di presentazione
  - i dati sono rappresentati in forma dimensionale – fatti e dimensioni
  - sia dati atomici che dati aggregati
  - progettazione per le prestazioni

Presentation server – modelli dimensionali

L'organizzazione dei dati nel data warehouse costituisce la pietra angolare dell'intero sistema DW/BI

- per ciascun processo di business dell'organizzazione di interesse, i dati sono rappresentati sulla base di *fatti* e *dimensioni*, mediante un *modello dimensionale* – anche chiamato *data mart*
- è anche necessario che i dati nei diversi data mart/modelli dimensionali siano organizzati in modo conforme (coerente)
  - questa conformità serve ad assicurare che i dati nei diversi data mart, relativi a processi diversi, possano essere correlati e integrati nell'ambito dell'intero data warehouse e nel corso del tempo
- per questo, viene adottata un'*architettura a bus del data warehouse*
  - nell'ambito dei diversi processi di business, i dati sono organizzati attorno a un insieme di dimensioni conformi e di fatti conformi

Back room



Back room

La **back room** – anche detta *area di preparazione dei dati* (*data staging area*) – è dove avviene l'elaborazione *ETL* (*Extract-Transform-Load*)

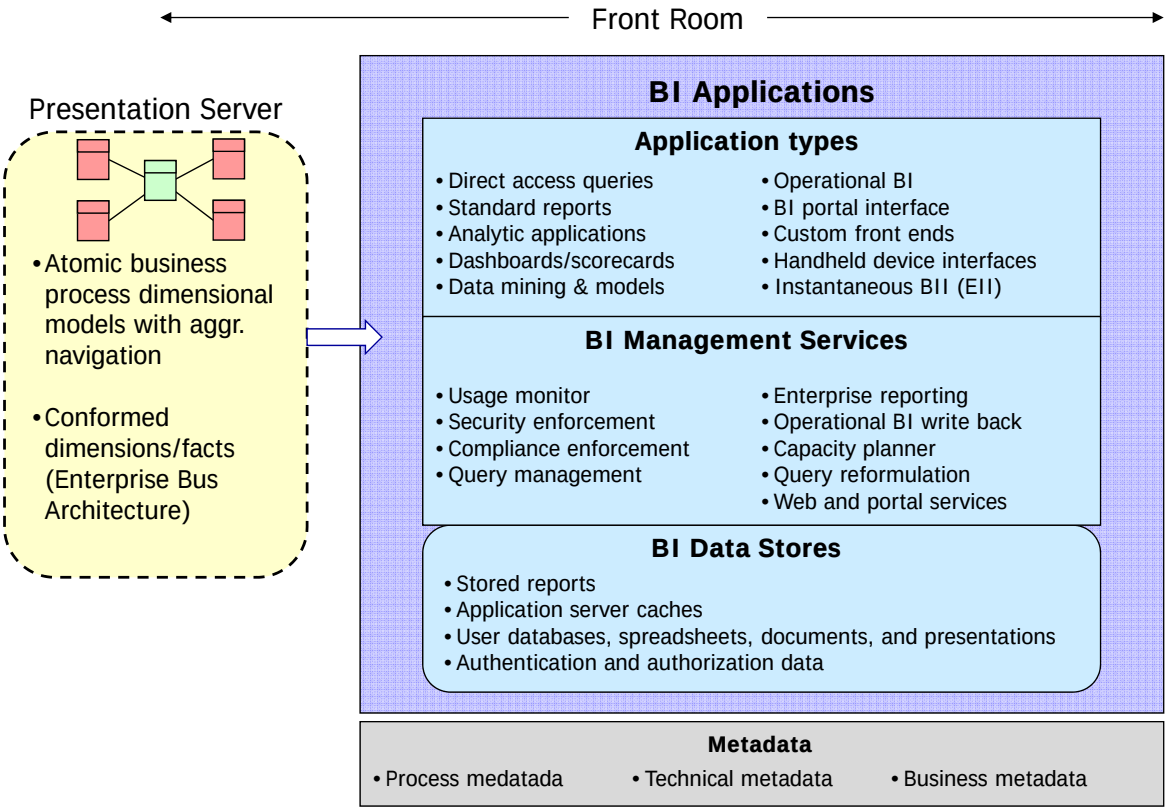
- l'interesse primario nella back room è fare in modo che i dati giusti di interesse si muovano da un punto A a un punto B, dopo un'appropriata trasformazione, nel momento appropriato
- quattro operazioni principali
  - *estrazione* dei dati dalle sorgenti informative
  - applicazione di un insieme di *trasformazioni di pulitura e conformazione*
  - *caricamento* nel DW – ovvero, nei server di presentazione
  - *gestione* dei processi ETL e dell'ambiente della back room

Back room – sottosistema ETL

La progettazione e lo sviluppo del *sottosistema ETL*

- è una delle attività più complesse con cui si deve confrontare il team di sviluppo
  - mediamente, il 70% dei rischi e delle sforzo di sviluppo in un progetto DW/BI sono relativi a questa attività
- ha a che fare con
  - estrazione dei dati dalle loro sorgenti informative
  - la qualità dei dati – pulizia, correlazione, conformazione, storicizzazione, ...
  - caricamento iniziale e caricamenti incrementali dei dati nel data warehouse

Front room



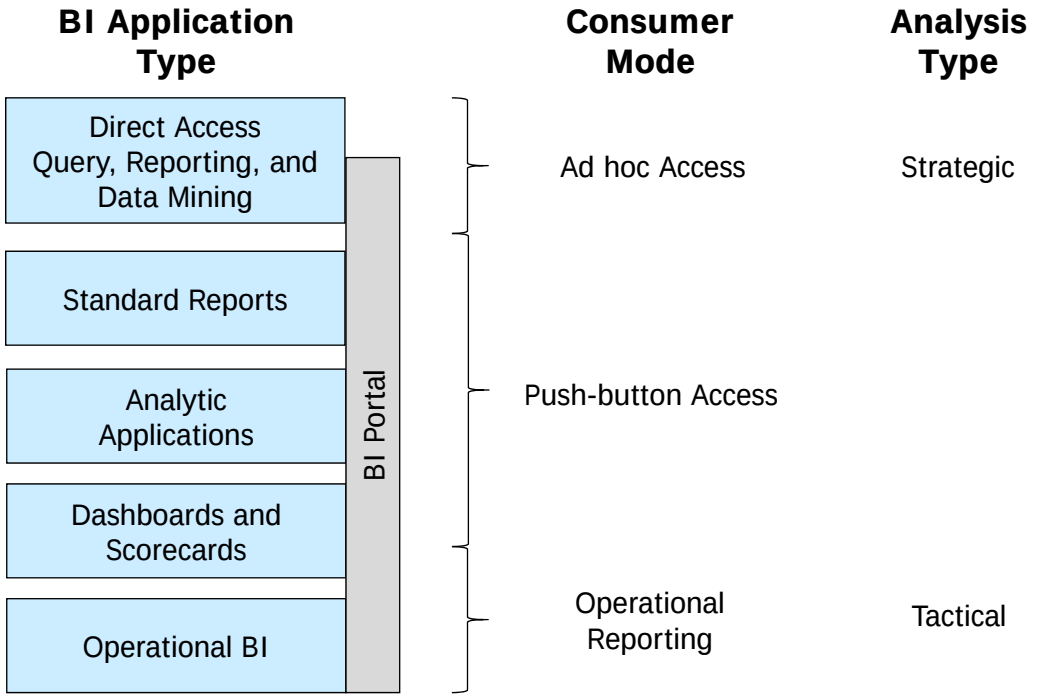
Front room

La **front room** è la faccia pubblica di un sistema DW/BI

- è composta da un insieme di *applicazioni di business intelligence* (*BI*) – utilizzate direttamente dagli utenti di business del DW/BI
- le applicazioni BI accedono direttamente ai dati del DW
- uno degli obiettivi primari di un sistema DW/BI è rendere le informazioni quanto più accessibili possibile
  - poiché i dati in un DW sono molto complessi, obiettivi (di usabilità) delle applicazioni BI comprendono
    - nascondere la complessità dei dati
    - aiutare gli utenti trovare quello che cercano
    - erogare i risultati nelle forme e nei formati richiesti
- inoltre, non tutti gli utenti hanno le stesse necessità di analisi
  - per questo, sono necessari diversi tipi di applicazioni BI
  - le applicazioni BI variano da semplici report standard parametrizzati a sistemi complessi di analisi dei dati

Front room – applicazioni BI

Tipi di applicazioni BI e modalità di consumo



Front room – ciclo di analisi di business

Un processo comune nelle analisi di business

- dal monitoraggio all'identificazione di un problema o di un'opportunità, dal determinare l'azione da intraprendere al monitorare il risultato di quell'azione
- attività di monitoraggio
  - spesso basata sull'uso di report standard, dashboard e scorecard – confrontando i risultati correnti con quelli attesi o di periodi precedenti
- identificazione di eccezioni
  - variazioni rispetto alle prestazioni normali possono indicare un problema oppure un'opportunità
- determinazione delle cause
  - identificazione delle cause delle eccezioni che sono state identificate – ad esempio, mediante analisi di tipo statistico o di data mining, o accesso ad ulteriori informazioni

Front room – ciclo di analisi di business

Un processo comune nelle analisi di business

- valutazione di alternative decisionali
  - vengono costruiti alcuni modelli che descrivono le relazioni causa-effetto dei fenomeni identificati
  - i dati nel data warehouse vengono utilizzati per validare i modelli costruiti – ma anche per prevedere l'effetto di alcune decisioni potenziali – sulla base di analisi statistiche, di tipo what-if e simulazioni
- intraprendere azioni e tracciare i risultati
  - idealmente, per chiudere il cerchio, le azioni raccomandate vengono "applicate" ai sistemi operazionali – inoltre, viene avviata la misurazione e il monitoraggio di queste azioni – per proseguire questo ciclo analitico

Ciclo di vita dimensionale

Il **ciclo di vita dimensionale** (**Business Dimensional Lifecycle** o **Kimball Lifecycle**) è una metodologia completa di progettazione e realizzazione di data warehouse

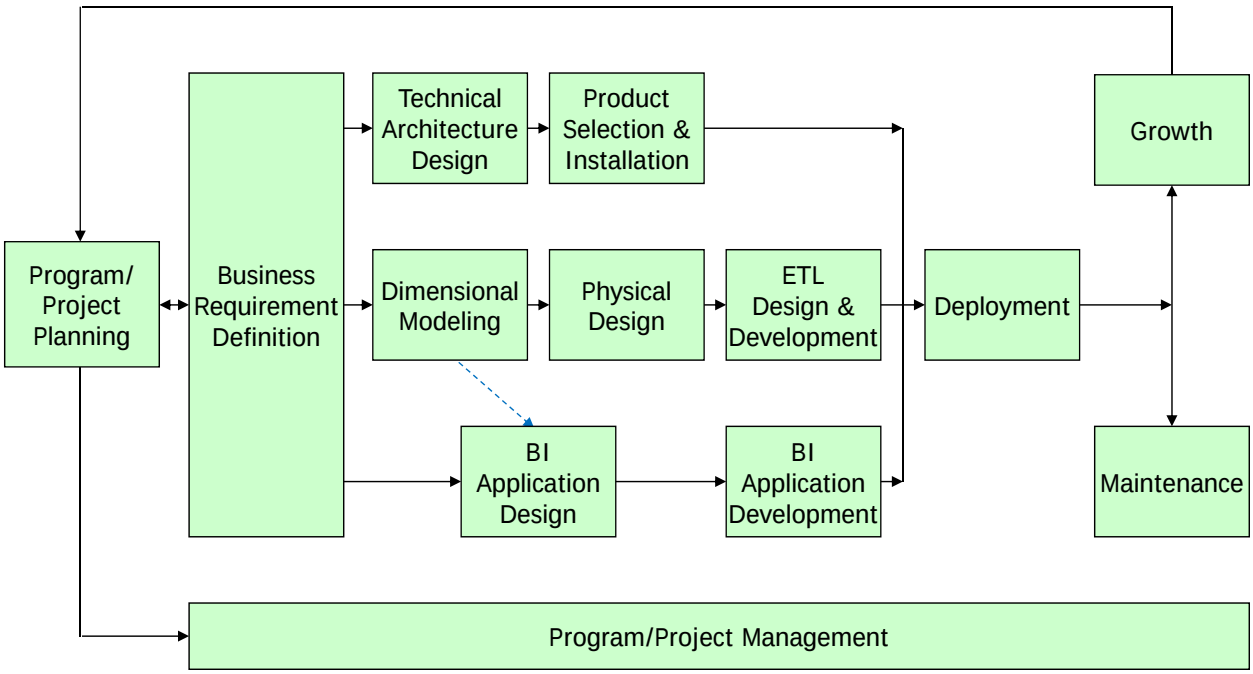
- fornisce il contesto di riferimento per la progettazione e realizzazione di data warehouse dimensionali
- il processo di sviluppo di un data warehouse è diverso da quello di altri sistemi software

Ciclo di vita dimensionale

Caratteristiche del ciclo di vita dimensionale

- un sistema DW/BI viene costruito a partire dagli utenti di business e dai loro bisogni effettivi – si lavora poi "all'indietro" attraverso i report, le applicazioni, le basi di dati e il software – fino agli strati più fisici del sistema
  - è un approccio fortemente guidato dal business e dall'utente – e non dalle tecnologie
- è un processo iterativo
  - l'intero sistema DW/BI ("programma") viene sviluppato in una serie di iterazioni ("progetti")

Ciclo di vita dimensionale



Ciclo di vita dimensionale

Elementi principali del ciclo di vita dimensionale

- pianificazione del progetto e del programma
- gestione del progetto e del programma
- raccolta e analisi dei requisiti
- progettazione del data warehouse – tre tracce concorrenti
  - progettazione dei dati
  - progettazione tecnologica
  - progettazione delle applicazioni BI
- installazione e avviamento
- manutenzione
- crescita