



The background features a blue grid pattern. Overlaid on this are various elements: binary code (0s and 1s) and hexadecimal strings (e.g., ED, 00, E, D1, F07, 0E801, 9D, 65EG, 6, 781, 65, 101E, 9CE, 10109A, 98E, 105, 3, 5) scattered across the top and middle. In the bottom right corner, there is a bar chart with approximately 20 vertical bars of varying heights, colored in a gradient from dark blue to light blue. The text 'DATA WAREHOUSE' is prominently displayed in the center in a bold, white, sans-serif font.

DATA WAREHOUSE

IL DATA WAREHOUSING

IL DATA WAREHOUSING

MODULO 6

INTRODUZIONE AL DATA WAREHOUSE

Il Ruolo Dell'informatica nei processi aziendali

- ❑ Il ruolo dei Sistemi Informatici è radicalmente cambiato dai primi anni '70 a oggi.
- ❑ I sistemi informatici si sono trasformati da semplici strumenti per migliorare l'efficienza dei processi a elementi centrali dell'organizzazione aziendale in grado di rivoluzionare la struttura dei processi aziendali



Tecnologie di supporto alla
gestione del Sistema Informativo

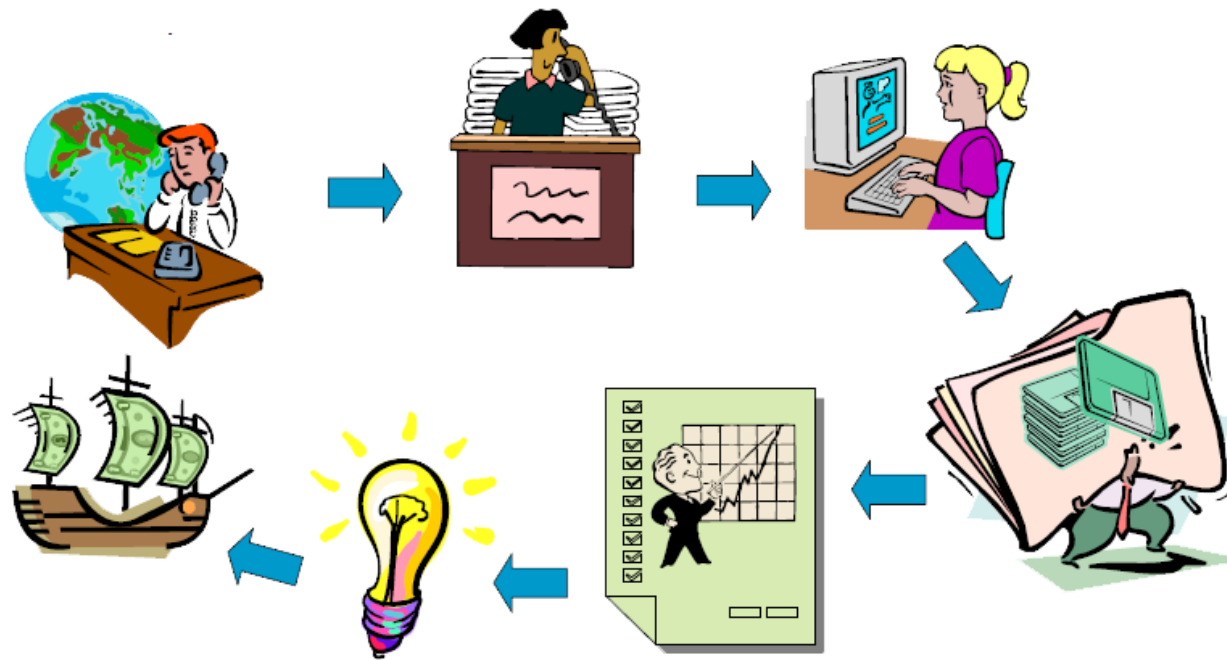
IL DUPLICE RUOLO
DELL'INFORMATICA



Disciplina organizzativa
che influenza i processi,
servizi e struttura aziendale

Dai dati alle informazioni

- ❑ Per ogni azienda è fondamentale poter disporre in maniera rapida e completa delle informazioni necessarie al processo decisionale: le indicazioni strategiche sono estrapolate principalmente dalla mole dei dati operazionali contenuti nei database aziendali, attraverso un procedimento di selezione e sintesi progressiva
- ❑ Uno scenario tipico è quello di una grande azienda, con numerose filiali, i cui dirigenti desiderano quantificare e valutare il contributo dato da ciascuna di esse al rendimento commerciale globale dell'impresa.



- ❑ Inevitabilmente in una organizzazione ci si ritrova a fare i conti con problematiche operative del tipo:
 - ❑ abbiamo montagne di dati ma non possiamo accedervi!
 - ❑ Questa frase esprime la frustrazione da parte di chi ha il ruolo e la competenza per decidere del futuro aziendale ma non possiede gli strumenti tecnici per ottenere, nella forma desiderata, i dati necessari.
 - ❑ come è possibile che persone che svolgono lo stesso ruolo presentino risultati sostanzialmente diversi?
 - ❑ In un contesto aziendale medio-grande sono tipicamente presenti più basi di dati, ciascuna relativa ad una diversa area del business, spesso memorizzate su piattaforme logico-fisiche differenti e non integrate dal punto di vista concettuale.
 - ❑ I risultati prodotti all'interno delle diverse aree saranno, allora, molto probabilmente inconsistenti tra loro.

- ❑ vogliamo selezionare, raggruppare e manipolare i dati in ogni modo possibile!
 - ❑ Il processo decisionale è difficilmente pianificabile a priori.
 - ❑ L'utente finale vorrebbe disporre di uno strumento sufficientemente amichevole e flessibile da consentirgli di condurre l'analisi in modo estemporaneo, lasciandosi guidare dalle informazioni via via ottenute per decidere sul momento quali nuove correlazioni ricercare.

- ❑ mostratemi solo ciò che è importante!
 - ❑ Esaminare i dati al massimo livello di dettaglio non solo è inutile ma controproducente per il processo decisionale perché non consente di focalizzare l'attenzione sulle informazioni veramente significative.

- ❑ tutti sanno che alcuni dati non sono corretti!
 - ❑ Questo è un altro punto dolente; una percentuale non trascurabile dei dati transazionali non è corretta o addirittura assente.
 - ❑ Evidentemente, basare il procedimento analitico su dati errati e incompleti non permette di raggiungere risultati validi.

- ❑ Da questo elenco di difficoltà e problemi possiamo facilmente estrarre un elenco di parole chiave che diventano fattori distintivi e requisiti indispensabili del processo di Data Warehousing, ossia del complesso di attività che consentono di trasformare i dati operazionali in conoscenza a supporto delle decisioni:
 - ❑ Accessibilità a utenti con conoscenze limitate di informatica e strutture dati;
 - ❑ Integrazione dei dati sulla base di un modello standard dell'impresa;
 - ❑ flessibilità di interrogazione per trarre il massimo vantaggio dal patrimonio informativo esistente;
 - ❑ Sintesi per permettere analisi mirate ed efficaci;
 - ❑ Rappresentazione multi-dimensionale per offrire all'utente una visione intuitiva ed efficacemente manipolabile delle informazioni;
 - ❑ correttezza e completezza dei dati integrati.

- ❑ I dati sono di vitale importanza per un'azienda, anche se ricavarne informazioni utili per il proprio business non è affatto semplice, infatti spesso si presentano in maniera caotica, con una struttura eterogenea dipendente dai sistemi sorgenti di provenienza
- ❑ Talvolta inoltre sono invalidati da una serie di errori di archiviazione con elementi duplicati, inseriti in modo non corretto, che influiscono sulla qualità finale del dato da analizzare con possibili conseguenze sull'analisi finale.
- ❑ Per questi motivi è necessario ricorrere alla definizione di un'architettura con cui vengono definiti gli elementi base di un data warehouse, ovvero uno o più schemi logici (schema dimensionale), algoritmi di integrazione e trasformazione dei dati (ETL) ed infine il sistema di archiviazione adeguato per offrire buone performance rispetto analisi OLAP (On-Line Analytical Processing) necessarie .

- ❑ Al centro del processo e di questa architettura si pone il data warehouse (DW); esso è un contenitore di dati che si fa garante dei requisiti esposti.
- ❑ Per «William H. Inmon» il data warehouse è una raccolta di dati che caratterizza dall' essere:
 - ❑ integrata (e consistente)
 - ❑ orientata ai soggetti di interesse
 - ❑ rappresentativa dell'evoluzione temporale e quindi dipendente dal tempo;
 - ❑ non volatile
- ❑ William H. Inmon (Bill Inmon) è considerato il padre fondatore del DW.

- ❑ Con Data Warehousing ci si riferisce al processo volto ad immagazzinare e organizzare le informazioni, ottimizzandone la struttura per un facile accesso e per supportare al meglio il processo decisionale aziendale.
- ❑ In questo processo i dati vengono estratti dalle varie fonti interne o esterne all'azienda, elaborati e ristrutturati all'interno dell'architettura logica del data warehouse per poterne estrarre le informazioni utili per il business aziendale.

Il Data Warehouse e la visione unificata

- ❑ Il DW si appoggia su più fonti di dati eterogenee:
 - ❑ dati estratti dall'ambiente di produzione, e quindi originariamente archiviati in basi di dati aziendali
 - ❑ provenienti da sistemi informativi esterni all'azienda.
- ❑ Di tutti questi dati il DW restituisce una visione unificata

- ❑ Perché il DWH è orientato a temi specifici dell'azienda (clienti, prodotti, ecc.) piuttosto che alle applicazioni o funzioni (quali ad esempio in un contesto bancario alle applicazioni transazionali).
- ❑ Nel DWH i dati vengono archiviati in modo che possano essere facilmente letti o elaborati dagli utenti cioè in modo da favorire la produzione di informazioni; viceversa i database operazionali sono organizzati intorno alle differenti applicazioni del dominio aziendale.

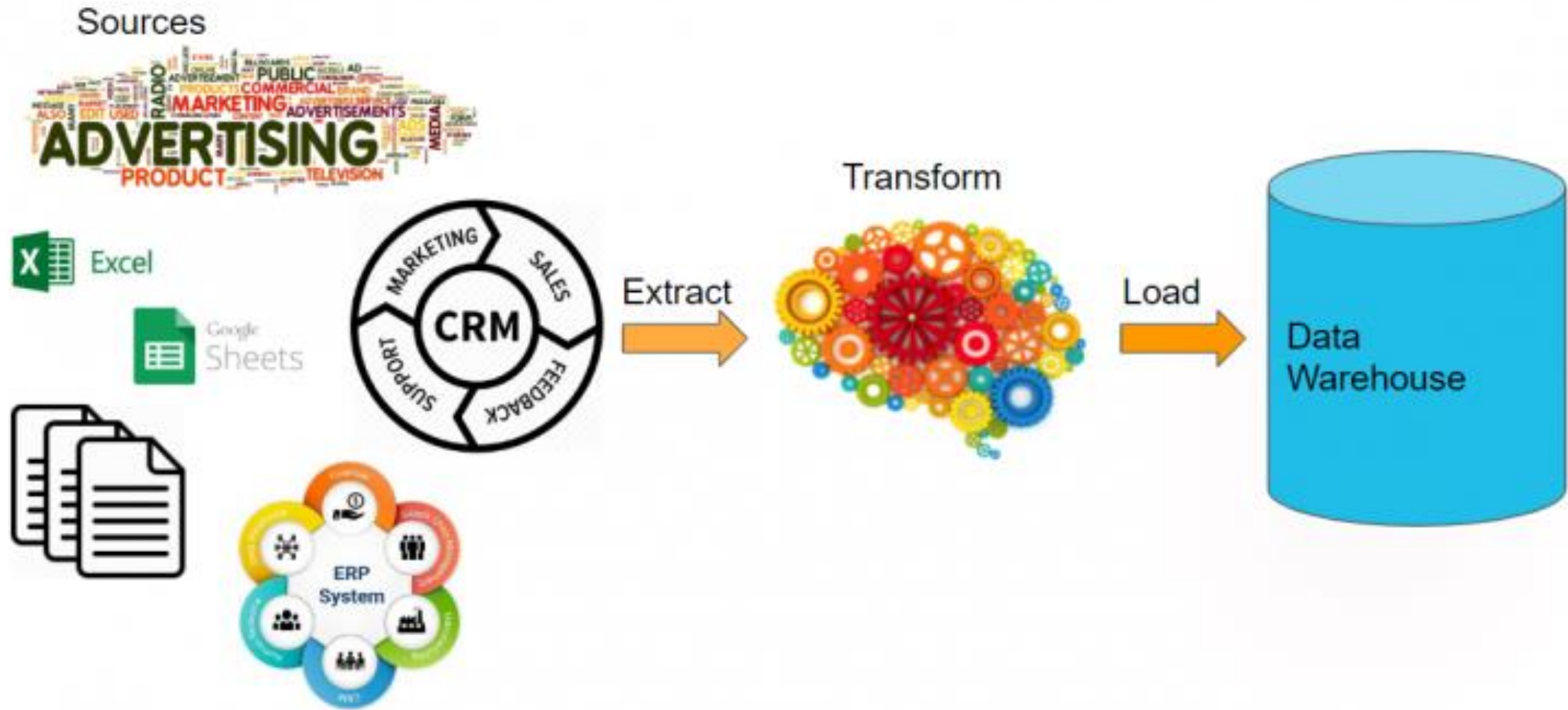
- ❑ L'integrazione è un requisito fondamentale del DWH in quanto in esso confluiscono dati provenienti da più fonti;
- ❑ L'obiettivo dell'integrazione può essere raggiunto percorrendo differenti strade: mediante l'utilizzo di metodi di codifica uniformi, mediante il perseguimento di una omogeneità sistematica di tutte le variabili, mediante l'utilizzo delle stesse unità di misura, ecc.
- ❑ La costruzione di un sistema di Data Warehouse non comporta l'inserimento di nuove informazioni bensì la riorganizzazione di quelle esistenti, e implica pertanto l'esistenza di un sistema informativo.

- ❑ I dati archiviati all'interno di un DW hanno un orizzonte temporale molto più esteso rispetto a quelli archiviati in un sistema operazionali.
- ❑ Nel DW sono contenute una serie di informazioni relative all'area di interesse che colgono la situazione relativa ad un dato fenomeno in un determinato intervallo temporale piuttosto esteso.
- ❑ Ciò, tuttavia, comporta che dati contenuti in un DW sono aggiornati fino ad una certa data, che nella maggior parte dei casi, è antecedente a quella in cui l'utente interroga il sistema.
- ❑ Situazione differente si manifesta in un database transazionale in cui i dati corrispondono sempre ad una situazione costantemente aggiornata che tuttavia non fornisce un quadro storico del fenomeno.

- ❑ Il dato viene caricato ed acceduto fuori linea cioè non può essere modificato dall'utente (l'accesso è in sola lettura).
- ❑ Si evitano le possibili anomalie dovute ad aggiornamenti e tanto meno si ricorre a strumenti complessi per gestire l'integrità referenziale o per bloccare record a cui possono accedere altri utenti in fase di aggiornamento.

- ❑ Un data warehouse viene quindi popolato tramite un processo ETL (Extraction Transformation and Load) in cui tutti i dati operazionali, esterni ed interni rispetto ai sistemi informatici aziendali, vengono:
 - ❑ Estratti: Attraverso opportuni connettori verso i sistemi di origine i dati vengono interrogati ed estratti per essere inseriti nel DWH;
 - ❑ Trasformati: è la fase principale del processo ETL, i dati devono essere prima filtrati e “puliti” rispetto a criteri di affidabilità e qualità, rimuovendo inconsistenze, errori di estrazione o archiviazione, ed infine “normalizzati” rispetto allo schema logico definito in fase di progettazione;
 - ❑ Caricati: la fase finale in cui i dati vengono caricati nel data sistema di archiviazione scelto, orientato all’ottimizzazione dei processi di lettura (es. Google BigQuery, Sistemi noSQL, etc...).

Popolamento di un Data Warehouse



Modello multidimensionale di un Data Warehouse

- ❑ È alla base della rappresentazione e l'interrogazione di dati nel data warehouse.
- ❑ A partire dalla definizione delle KPI (Key Performance Indicator), ovvero indicatori sintetici per esprimere in modo sintetico l'andamento di un processo aziendale, vengono identificate le metriche di composizione delle stesse e le possibili dimensioni di visualizzazione.
- ❑ Le dimensioni definiscono quindi la prospettiva da cui visualizzare una determinata metrica
- ❑ le metriche invece costituiscono il valore da misurare, sono quindi un valore numerico (es. fatturato, numero di vendite, importo venduto, etc...).
- ❑ Il nome modello dimensionale deriva quindi dalla definizione di n-dimensioni di analisi della metrica di interesse, rappresentabili tramite un ipercubo, dove il singolo elemento costituisce il fatto (es. vendita di un prodotto) mentre la prospettiva è come visualizzarlo (es. quando, dove, cosa, chi, etc...)

- ❑ Una volta progettato e popolato il data warehouse la fase finale consiste nel progettare dashboard utili al management per visualizzare nel modo più efficace possibile le informazioni raccolte.
- ❑ In particolare è possibile visualizzare i dati in:
 - ❑ Reportistica statica, ovvero estrazioni schedate dei dati elaborati;
 - ❑ Reportistica dinamica (OLAP), analisi dinamiche attraverso strumenti grafici che permettono la “navigazione” del dato raccolto, offrendo funzioni di drill-down, drill-up, drill-through, scomponendo la metrica per le varie dimensioni di analisi;
 - ❑ Dashboards che tramite visualizzazioni grafiche (diagrammi, tabelle, etc...) facilitano la lettura del dato.

- ❑ In conclusione l'era dei big data ha segnato una trasformazione senza precedenti per le imprese, costringendole a confrontarsi con una sfida evolutiva che le ha portate ad adattarsi ai più recenti strumenti di raccolta, integrazione e analisi dei dati, alle nuove ricerche di mercato, ai nuovissimi modelli di web marketing, nonchè ad un inedito rapporto con la clientela.
- ❑ Strumenti come il data mining, il data warehousing e la data visualization e tutto ciò che compone il complesso mondo della Business Intelligence, sono diventati, nel corso di pochissimi anni, elementi chiave per le aziende che vogliano risultare efficaci nel mercato odierno e in grado di sfidare al meglio la concorrenza.

MODULO 7

DATABASE vs DATA WAREHOUSE

Data Warehouse = Creare Conoscenza

- ❑ Data Warehouse = Trasformare il dato in informazione
- ❑ Data Warehouse = Consolidazione dei dati operazionali da sistemi legacy diversi in un unico deposito informativo.

Perché accontentarsi di un Data Base quando si può avere un Data Warehouse?

- ❑ Mentre i dati operazionali coprono un arco temporale di solito piuttosto limitato, poiché la maggior parte delle transazioni coinvolge i dati più recenti, il DW deve permettere analisi che spaziano sulla prospettiva di alcuni anni.
- ❑ Per questo motivo, il DW è aggiornato a intervalli regolari a partire dai dati operazionali ed è in crescita continua.
- ❑ Volendo fare un paragone possiamo supporre che, a intervalli regolari, venga scattata una fotografia istantanea dei dati operazionali.
- ❑ La progressione delle fotografie scattate viene immagazzinata nel DW, dove genera un film che documenta la situazione aziendale da un istante zero fino al tempo attuale.
- ❑ Proprio il fatto che, in linea di principio, non vengono mai eliminati dati dal DW e che gli aggiornamenti siano tipicamente eseguiti “a freddo”, ossia quando il DW è fuori linea, fa sì che un DW possa essere fondamentalmente considerato come un **database a sola lettura**

Database Vs DW

Riferimento	DB Transazionali	DW
Utenti	Migliaia	Centinaia
Carico di lavoro	transaction - intensive, Transazioni predefinite	query - intensive, Interrogazioni ed analisi ad hoc
Accesso	Centinaia di record in lettura e scrittura	Milioni di record per lo più in lettura
Scopo	Dipende dall'applicazione; i sistemi transazionali sono definiti per un limitato dominio applicativo che si riferisce a una specifica applicazione	Supporto alle decisioni; i progetti di DW forniscono un'infrastruttura di supporto ai sistemi di supporto alle decisioni con caratteristiche di scalabilità di ampliamento e flessibilità
Dati	Elementari, dati aggregati per attività o per processo	Di sintesi, dati integrati per soggetto
Integrazione dei dati	Per applicazione	Per soggetto
Qualità	In termini di integrità	In termini di consistenza
Copertura temporale	Solo dati correnti	Dati correnti e storici
Aggiornamenti	Continui	periodici
Modello	Normalizzato; il modello è conforme alle esigenze che derivano dalle transazioni da supportare	Denormalizzato, Multidimensionale; il modello è conforme alle dimensioni del soggetto
Ottimizzazione	per accessi OLTP su una frazione del DB	Per accessi OLAP su gran parte del DB
Sviluppo	A cascata; sistemi OLTP sviluppati seguendo i requisiti del sistema esplicitati dagli utenti e il metodo di sviluppo a cascata	iterativo; i criteri di sviluppo rispondono a Sviluppo principi evolutivi e iterativi
Database	dati aggiornati continuamente e quindi volatili	dati aggiornati ad intervalli predefiniti, perciò non-volatile
Sponsorship	i sistemi transazionali tendono ad essere sponsorizzati seguendo un processo che consente di individuare il responsabile che individua a sua volta anche le gerarchie organizzative	un progetto di DW richiede una forte sponsorizzazione a causa dell'ampiezza organizzativa dello stesso

Database Vs DW

Riferimento	DB Transazionali	DW
Tipo Di Utente	Impiegato	Dirigente
Funzione	Operazioni giornaliere	Supporto alle decisioni
Progettazione	Orientata all'applicazione	Orientata al dato
Dati	Correnti, aggiornati, dettagliati, relazionali, omogenei	Sorici,m aggregati, multidimensionali, eterogenei
Uso	Ripetitivo	Casuale
Accesso	Read-Write, indicizzato	Read, sequenziale
Unità di lavoro	Transazione Breve	Interrogazione complessa
Dimensione media	100MB-10GB	100GB-10TB
Metrica	Throughput (Capacità di trasmissione)	Tempi di risposta

- ❑ Per i database operazionali, le interrogazioni eseguono transazioni che in genere leggono e scrivono un ridotto numero di record da diverse tabelle legate da semplici relazioni: per esempio, si ricercano i dati di un cliente per inserire un suo nuovo ordine. Questo tipo di elaborazione viene comunemente detto On-Line Transactional Processing (OLTP).
- ❑ Al contrario, il tipo di elaborazione per cui nascono i DW viene detto On-Line Analytical Processing (OLAP), ed è caratterizzato da un'analisi dinamica e multidimensionale che richiede la scansione di un'enorme quantità di record per calcolare un insieme di dati numerici di sintesi che quantificano le prestazioni dell'azienda.
- ❑ Perché dati separati?
 - ❑ Prestazioni: Ricerche complesse riducono le prestazioni delle transazioni operative; metodi di accesso diversi a livello fisico.
 - ❑ Gestione dei dati: informazioni mancanti (storico); consolidamento dei dati; qualità dei dati problema di inconsistenze).

- ❑ I dati disponibili ad un'organizzazione sono spesso memorizzati su sorgenti informative pre-esistenti e differenziate
- ❑ Collezioni eterogenee di ambienti hardware e software
 - ❑ Modelli logici differenti
 - ❑ Linguaggi di definizione e di manipolazione dei dati differenti
 - ❑ Meccanismi di gestione della concorrenza e delle transazioni eterogenei

- ☐ È necessario garantire la massima trasparenza per gli utenti finali
- ☐ È necessario garantire la massima autonomia alle basi di dati coinvolte
- ☐ Trasparenza ed autonomia sono molto spesso in conflitto
- ☐ Al fine di raggiungere una elevata efficienza è necessario minimizzare le interazioni tra i componenti
- ☐ Cooperazione: garantire la massima trasparenza, autonomia ed efficienza

- ☐ Gli schemi delle basi di dati da integrare possono essere differenti
 - ☐ Automobili(numTelaio, modello, colore, cambioAutomatico, lettoreCD, ...)
 - ☐ Autovetture(nTelaio, modello, colore) Optional(nTelaio, optional)
- ☐ Nomi diversi possono essere utilizzati per rappresentare lo stesso concetto (sinonimie)
- ☐ Nomi uguali possono essere utilizzati per rappresentare concetti diversi (omonimie)
- ☐ Differenze nel tipo di dati
- ☐ Differenze nei valori
- ☐ Differenze nella semantica
- ☐ Valori mancanti

- ❑ La quantità di dati disponibili rende la loro interpretazione ai fini della gestione dei processi organizzativi spesso complessa, se non impossibile, con strumenti tradizionali
- ❑ Spesso i dati non vengono utilizzati
- ❑ Decisioni importanti vengono prese senza sfruttare tutte le informazioni disponibili
- ❑ È necessario cercare di colmare il gap tra dati e informazioni estraendo informazioni utili dai dati disponibili

- ❑ Database Distribuiti
 - ❑ vi è un unico DBMS ma i dati sono distribuiti su più siti
- ❑ Sistemi Informativi Federati
 - ❑ vi sono più sorgenti indipendenti; ciascuna può richiedere alle altre informazioni
- ❑ Sistemi Informativi Cooperativi
 - ❑ le sorgenti sono indipendenti; al di sopra di esse vi è un componente software che supporta un database virtuale che l'utente può interrogare come se lo stesso fosse materializzato
- ❑ Data Warehousing
 - ❑ le copie dei dati provenienti da sorgenti differenti sono memorizzate in un singolo database integrato, denominato Data Warehouse, dedicato al supporto delle decisioni

- ☐ Il DBMS è unico; i dati sono però memorizzati (ed, eventualmente, replicati) su più siti
- ☐ In questo modo si massimizza la trasparenza ma si annulla l'autonomia
- ☐ Possibili forme di distribuzione dei dati nei siti:
 - ☐ Replicazione
 - ☐ Frammentazione
 - ☐ Replicazione e Frammentazione

- ❑ L'integrazione completa dei sistemi esistenti in un database omogeneo è spesso difficile per:
 - ❑ difficoltà tecniche: costo di conversione alto
 - ❑ difficoltà organizzative: i DBMS esistenti appartengono a diversi dipartimenti o organizzazioni
- ❑ Generalmente, nell'effettuare l'informatizzazione di un'organizzazione complessa, emerge sempre la necessità di garantire un alto grado di autonomia ai singoli dipartimenti
- ❑ Negli ultimi anni questo trend si è via via sempre più consolidato

- ❑ Connessioni uno-ad-uno tra tutte le coppie di sorgenti informative
- ❑ IS-j può interrogare IS-i utilizzando il linguaggio e le modalità che quest'ultimo è in grado di comprendere
- ❑ Problema: con n sorgenti informative è necessario scrivere $n(n-1)$ Sistemi Federati porzioni di codici di traduzione
- ❑ Tale architettura è adeguata quando le comunicazioni nel sistema avvengono prevalentemente tra singole coppie di sorgenti o quando il numero delle sorgenti coinvolte è piccolo

- ❑ Sistemi cooperativi: risolvono solo il problema dell'accesso trasparente all'informazione
- ❑ DWs: danno valore aggiunto all'informazione disponibile, fornendo ulteriori servizi informativi che vanno al di là del mero accesso trasparente all'informazione

MODULO 8

Cosa non è il Data Warehouse
&
Consigli

Un prodotto che può essere acquistato e installato.

- ❑ Il Data Warehouse è un SISTEMA INFORMATIVO che va sviluppato da zero utilizzando, è vero, tecnologia software e hardware specifica, ma una volta acquistato il prodotto si è fatto sì e no il 5% del lavoro, senza contare che molto spesso la scelta si rileva sbagliata o sottodimensionata perché non è stata fatta a fronte di un'attenta,

Una soluzione di business in sé e per sé.

- ❑ Il Data Warehouse non va progettato e creato in quanto tale.
- ❑ Un'azienda non affronta un progetto di «data warehousing» in quanto tale, ma dovrebbe decidere in maniera seria e ragionata di risolvere una specifica esigenza di business (per esempio il Customer Care) utilizzando la tecnologia del data warehouse.
- ❑ Se un data warehouse parte con presupposti diversi da questi è destinato a fallire.

Una 'pezza' ai problemi dei sistemi transazionali

- ☐ Il data warehouse è un'ottima risposta a esigenze di analisi e supporto decisionale, funzionalità supportate poco o male dai sistemi transazionali. Punto.
- ☐ Non deve essere utilizzato per risolvere altri problemi che non è deputato a risolvere.
- ☐ Alcuni esempi:
 - ☐ Calcolo dei costi di produzione;
 - ☐ Stampa delle etichette per il mailing;
 - ☐ Ribaltamenti e 'spalmature' della contabilità analitica;
 - ☐ Stampa del conto economico aziendale (civilistico);
 - ☐ Contact management;
 - ☐ Data entry manuale di informazioni;

- ❑ Se qualcuno vi dice che intende usare il data warehouse per una o più di queste cose fermatelo prima che si faccia del male: spenderà tantissimi soldi e otterrà risultati scarsi.
- ❑ In realtà molto spesso chi crea confusione su cosa sia e come si fa un data warehouse sono i software vendor e le società di consulenza, gli uni nel tentativo di legare i propri prodotti a slogan e concetti che nel mercato «tirano», le seconde nel costante tentativo di inventare qualcosa di nuovo da proporre ai clienti.
- ❑ Ecco che così un report writer qualsiasi diventa «un prodotto OLAP» ed un database su PC usato per stampare conti economici per prodotto diventa «il data warehouse dell'analisi della redditività».
- ❑ Costruire un data warehouse è un'impresa lunga e complessa dove nulla va banalizzato e dove le competenze di tipo tecnologico contano esattamente quanto quelle funzionali

- ❑ Non esiste una metodologia unica, consolidata e sempre valida per la realizzazione di un DW.
- ❑ Il motivo è semplice: realizzare sistemi di Decision Support non è proprio una scienza esatta, è piuttosto un percorso lungo il quale si effettuano spesso prove e s’impara dagli errori commessi.
- ❑ Arrivare ad DW “perfetto” (ammesso che esista) per una specifica azienda (o istituzione) è un obiettivo che si raggiunge per approssimazioni successive.

Think Big, Start Small

- ❑ Non bisogna strafare, ma neanche avere un'ottica di breve periodo.
- ❑ Per essere pratici: non bisogna iniziare un progetto di DW con l'obiettivo di realizzare semplicemente l'architettura e finire qui;
- ❑ Il DW è figlio di un processo adattivo ed evolutivo che si trasforma nel tempo.
- ❑ Cosa significa? Significa che bisogna iniziare da aree analitiche relativamente semplici e conosciute in termini di dati e problematiche e costruire una serie di “storie di successo” che convincano l'organizzazione che il data warehouse è la strada giusta.
- ❑ Però è sbagliato compiere scelte (soprattutto da un punto di vista tecnologico) che non siano «scalabile», cioè che richiedano di essere cambiate nel momento in cui la complessità ed il volume dei dati crescono.
- ❑ In sintesi quindi gli obiettivi devono essere a breve periodo, la visione del progetto deve invece guardare necessariamente lontano.

- ❑ Alcuni sostengono che un progetto di DW si può fare solo e unicamente con gente che vive per costruire architetture di warehousing.
- ❑ Ciò perchè Il ventaglio delle tematiche da affrontare in tale tipologia di progetto è talmente vasta che per costruire un esperto di data warehousing ci vogliono anni, non mesi.
- ❑ Se a ciò si aggiungono le continue e costanti evoluzioni tecnologiche (Teorie, tools etc..) risulta chiaro che chi affronta un progetto di data warehousing deve considerare di mettere in campo quante più conoscenze ed esperienze possibili in vari ambiti di competenza.

Partire dall'esigenza, non dal prodotto.

- ☐ Come già detto, il DW non si compra.
- ☐ Diffidate di chiunque sostenga il contrario, pacchettizzare una piattaforma di decision support non è possibile ed in ogni caso ha delle controindicazioni evidenti.
- ☐ Nello stesso tempo è scorretto selezionare un prodotto e poi costruire il DW con tale prodotto.
- ☐ La software selection può derivare unicamente da un'analisi attenta e seria delle esigenze funzionali da soddisfare e dell'infrastruttura tecnica e tecnologica di cui il sistema di DW sarà una componente.
- ☐ La realizzazione di un DW deve assolutamente prendere spunto da un'analisi degli aspetti di business.
- ☐ Un DW non focalizzato su uno o più specifici aspetti di business sarà un DW senza utenti e quindi destinato a morire di inedia e solitudine.

Non banalizzare gli aspetti tecnologici

- ❑ Non bisogna però nemmeno eccedere nella direzione esattamente opposta.
- ❑ La realizzazione di un DW pone tali e tanti problemi di capacity planning e di definizione architetturale che un'attenzione meno che ossessiva a questo aspetto può decretare il fallimento del sistema.
- ❑ Non sono accettabili tempi di attesi lunghi, né interfacce poco user friendly, né continue failure di sistema dovute a mancanza di memoria, CPU,etc.

Attenzione alla qualità dei dati

- ❑ Secondo alcune ricerche circa il 50% degli insuccessi nei progetti di DW sono dovuti a problemi di scarsa qualità dei dati alimentati.
- ❑ E' opportuno, prima di partire con l'implementazione del DW, effettuare una verifica sulla qualità dei dati alimentati, onde evitare brutte sorprese in seguito.
- ❑ Per la verità molti problemi di data quality vengono scoperti durante il progetto ma un'analisi preliminare ben fatta permette di individuare eventuali aree particolarmente critiche o delle problematiche tipiche dei sistemi informativi alimentanti
- ❑ In definitiva bisogna cercare di abbattere il rischio derivante da una sovrabbondanza di dati mal interpretanti; un errore in fase di filtro, rincociliazione e aggregazione dei dati porterà tutta la sovrastruttura di BI a «decisioni errate»
- ❑ Gettare fondamenta solide significa erigere un edificio senza paura che crolli.

Lasciare perdere il report layout

- ❑ L'approccio tradizionale alla progettazione dei sistemi di supporto decisionale prevede che si parta dai fabbisogni informativi, cioè dalle informazioni di cui gli utenti del sistema vorranno la disponibilità.
- ❑ Per questo motivo molti, invece di preoccuparsi di analizzare il modello dimensionale di business dell'azienda, realizzano improbabili e artistici report layout andando in giro a chiedere: "E' questa l'informazione che vuoi vedere?".
- ❑ Questo tipo di lavoro ha poco senso ed è pericoloso.
- ❑ Ha poco senso in quanto l'utilizzo del DW avviene mediante dei report writer (mediante i quali l'utente stesso può disegnarsi i report) o attraverso degli OLAP browser (in cui il concetto di report non esiste).
- ❑ E' pericoloso perché può capitare che i report finali siano diversi da quelli presentati e inoltre perché da un punto di vista di layout qualsiasi foglio elettronico consente destrutturate perversioni artistiche, inimmaginabili per un OLAP viewer, tradizionalmente basato su uno schema dati di tipo rigorosamente multidimensionale

MODULO 9

Inmon vs Kimball

Chi sono Inmon e Kimball ?

☐ William Inmon

- ☐ è conosciuto come “il padre del Data Warehousing” dato che nel 1991 ha coniato il termine “Data Warehouse”.
- ☐ Inmon ha definito un modello per supportare la “single version of the truth”

☐ Ralph Kimball

- ☐ è conosciuto come “il padre della Business Intelligence” per aver coniato il termine “Data Mart”
- ☐ Kimball ha concettualizzato gli star-schema e gli schemi snowflakes supportando lo sviluppo di tool che utilizzano gerarchie dimensionali.

L'approccio di Inmon

- ❑ Inmon identifica la necessità di integrare dati provenienti da diverse sorgenti in un repository centralizzato con la finalità di avere supporto alle decisioni.
- ❑ Secondo Inmon, i dati devono essere organizzati in strutture subject-oriented, integrated, time-variant e nonvolatile.
- ❑ I dati devono essere ottenibili ad un livello di granularità fine (mediante operatori di drill-down) o più grossolana (mediante operatori di roll-up).
- ❑ Inmon segue un approccio top-down.
 - ❑ I dati vengono estratti dalle sorgenti
 - ❑ I dati vengono caricati nella staging area
 - ❑ Viene effettuato il data cleansing
 - ❑ I dati vengono trasformati, integrati e consolidati
 - ❑ I dati vengono caricati nel Datawarehouse

- ❑ Kimball disegna il datawarehouse come la connessione di data marts mediante una struttura a bus.

datawarehouse virtuale = unione di data marts

- ❑ L'approccio di Kimball è reso possibile dall'utilizzo di dimensioni conformi nei data marts.
- ❑ Kimball segue un approccio bottom-up.
 - ❑ I dati vengono estratti dalle sorgenti
 - ❑ I dati vengono caricati, consolidati e verificati nella staging area
 - ❑ I dati vengono caricati ed integrati con altri dati nel data store. Il data store contiene la copia aggiornata dei dati.
 - ❑ I dati vengono aggregati e trasferiti nei data marts.

Inmon vs Kimball: le differenze

☐ Visione di Kimball:

- ☐ non riconosce la necessità di un datawarehouse centralizzato.
- ☐ supporta la trasformazione di dati grezzi in informazioni per il supporto alle decisioni.
- ☐ La trasformazione avviene nella staging area, prima del loro caricamento nei data marts.

☐ Visione di Inmon:

- ☐ riconosce il datawarehouse centralizzato come l'insieme dei dati organizzati in maniera significativa per l'enterprise.
- ☐ Data marts dipendenti rappresentano diverse "viste" sugli stessi dati.

- ❑ Kimball e Inmon concordano che il successo del datawarehouse e/o dei data marts dipende dalle necessità di business.

La raccolta dei requisiti è una fase cruciale

- ❑ Entrambi considerano rilevante la validazione dei data marts da parte degli utenti finali.
- ❑ Entrambi adottano il concetto di staging.

Approccio top-down

- ❑ Analizza i bisogni globali dell'intera azienda e pianifica lo sviluppo del DW per poi progettarlo e realizzarlo nella sua interezza
- ❑ Promette ottimi risultati poiché si basa su una visione globale dell'obiettivo e garantisce in linea di principio di produrre un DW consistente e ben integrato
- ❑ Il preventivo di costi onerosi a fronte di lunghi tempi di realizzazione scoraggia la direzione dall'intraprendere il progetto
- ❑ Affrontare contemporaneamente l'analisi e la riconciliazione di tutte le sorgenti di interesse è estremamente complesso
- ❑ Riuscire a prevedere a priori nel dettaglio le esigenze delle diverse aree aziendali impegnate è pressoché impossibile, e il processo di analisi rischia di subire una paralisi
- ❑ Il fatto di non prevedere la consegna a breve termine di un prototipo non permette agli utenti di verificare l'utilità del progetto e ne fa scemare l'interesse e la fiducia

Approccio bottom-up

- ☐ Il DW viene costruito in modo incrementale, assemblando iterativamente più data mart, ciascuno dei quali incentrato su un insieme di fatti collegati a uno specifico settore aziendale e di interesse per una certa categoria di utenti
- ☐ Determina risultati concreti in tempi brevi
- ☐ Non richiede elevati investimenti finanziari
- ☐ Permette di studiare solo le problematiche relative al data mart in oggetto
- ☐ Fornisce alla dirigenza aziendale un riscontro immediato sull'effettiva utilità del sistema in via di realizzazione
- ☐ Mantiene costantemente elevata l'attenzione sul progetto
- ☐ Determina una visione parziale del dominio di interesse

MODULO 10

SCELTA DELL'ARCHITETTURA

- ❑ Molte organizzazioni mancano della necessaria esperienza e capacità per affrontare con successo le sfide implicite nei progetti di data warehousing
- ❑ Uno dei fattori che maggiormente minaccia la riuscita dei progetti è la mancata adozione di una approccio metodologico, che minimizza i rischi di insuccesso essendo basato su un'analisi costruttiva degli errori commessi

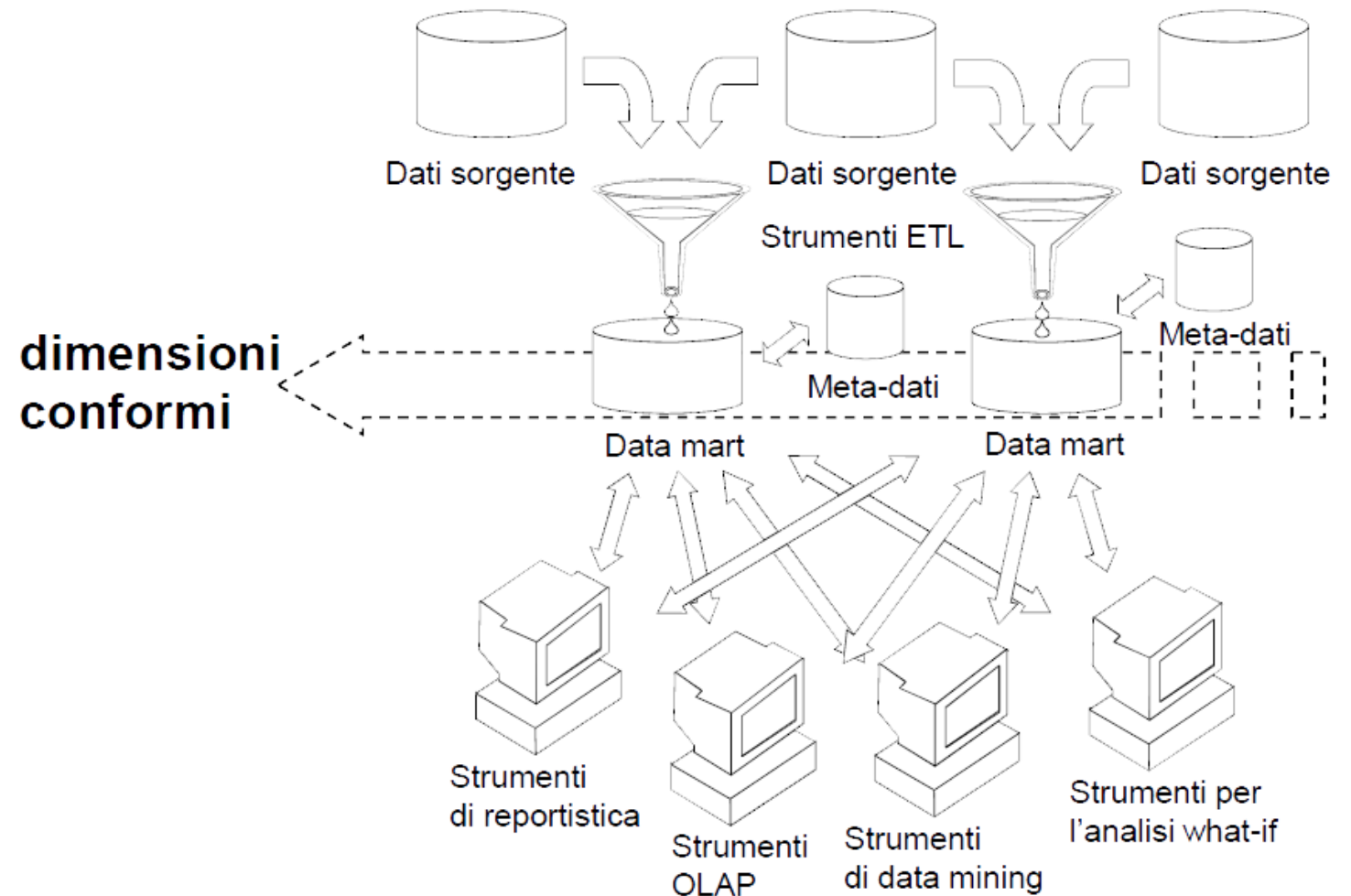
- ❑ Fattori di Rischio:
 - ❑ Rischi legati alla gestione del progetto
 - ❑ Rischi legati alle tecnologie
 - ❑ Rischi legati ai dati e alla progettazione
 - ❑ Rischi legati all'organizzazione
- ❑ Il rischio di ottenere un risultato insoddisfacente nei progetti di data warehousing è particolarmente alto a causa delle elevatissime aspettative degli utenti
- ❑ Nella cultura aziendale contemporanea è infatti diffusissima la credenza che attribuisce al data warehousing il ruolo di panacea
- ❑ In realtà una larga parte della responsabilità della riuscita del progetto ricade sulla qualità dei dati sorgente e sulla lungimiranza, disponibilità e dinamismo del personale dell'azienda

- ❑ Obiettivi:
 - ❑ Soddisfare le esigenze di business manifestate
 - ❑ Garantire
 - Rapido ritorno dell'investimento (ROI)
 - Flessibilità / apertura a nuove esigenze
 - Minimo impatto sull'ambiente di produzione
 - Congruenza dei risultati delle interrogazioni
 - Integrazione dei dati
 - Scalabilità
- ❑ Due architetture di riferimento
 - ❑ La “Data Mart Bus Architecture”
 - ❑ La “Hub and Spoke Architecture”

- ❑ Come accennato un Data Mart è un sottoinsieme o un'aggregazione dei dati presenti nel DW primario, contenente l'insieme delle informazioni rilevanti per una particolare area del business, una particolare divisione dell'azienda, una particolare categoria di soggetti.
- ❑ I data mart alimentati dal DW primario sono detti dipendenti. Per i sistemi collocati all'interno di realtà aziendali medio-grandi essi sono utili:
 - ❑ come blocchi costruttivi durante la realizzazione incrementale del DW;
 - ❑ in quanto delineano i contorni delle informazioni necessarie a un particolare tipo di utenti per le loro interrogazioni;
 - ❑ poiché, essendo di dimensioni inferiori al DW primario, permettono di raggiungere prestazioni migliori
- ❑ In alcuni contesti si preferisce adottare data mart alimentati direttamente dalle sorgenti, detti indipendenti
 - ❑ L'assenza di un DW primario snellisce le fasi progettuali, ma determina uno schema complesso di accessi ai dati e ingenera il rischio di incongruenze tra i data mart

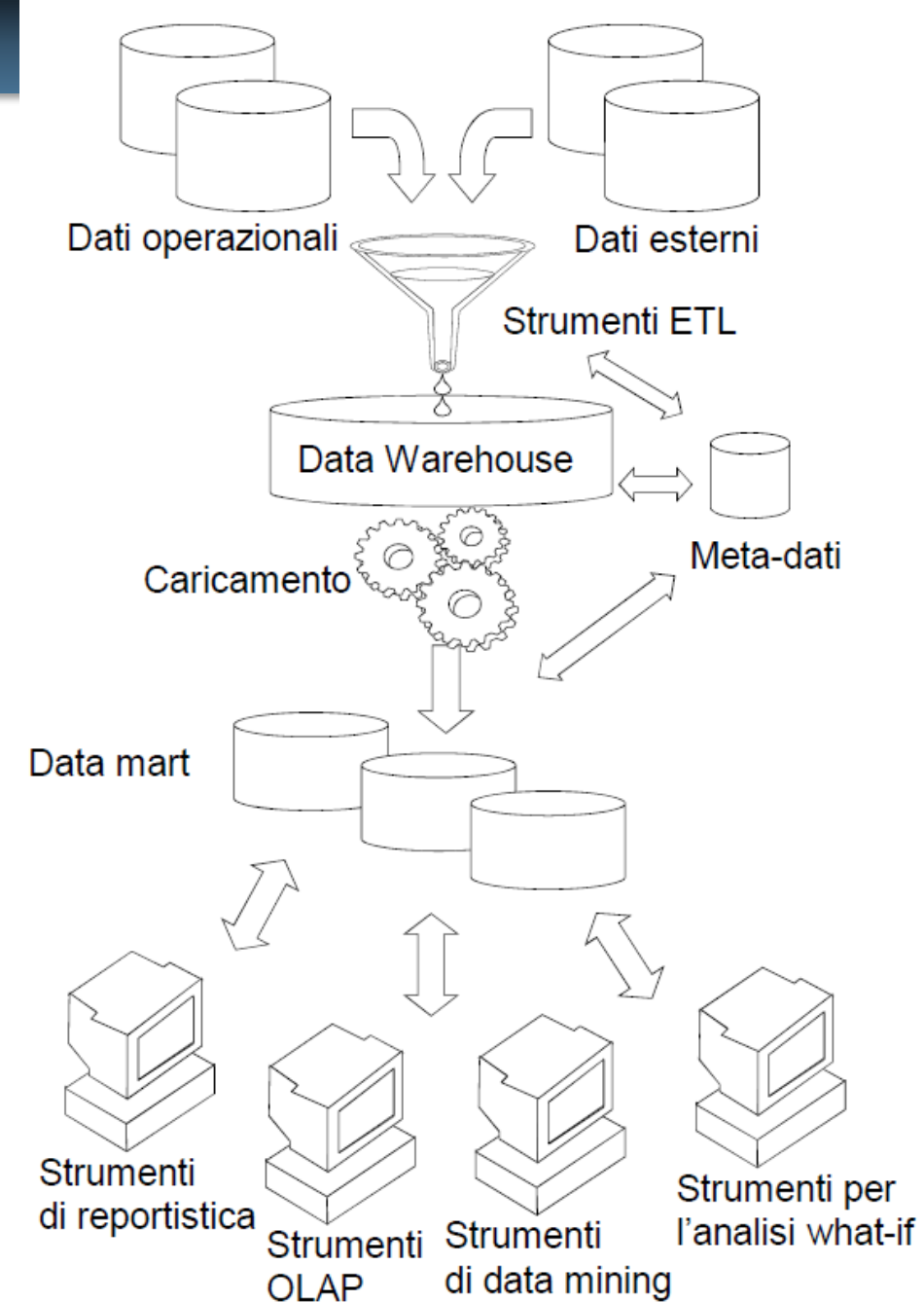
Architettura Data Mart Bus

- ❑ Data Mart tra loro logicamente integrati da dimensioni comuni
- ❑ «Enterprise view»
- ❑ Approccio consigliato da Kimball



Architettura Hub-and-spoke

- ❑ Letteralmente “mozzo e raggi”
- ❑ Una delle architetture più usate in contesti medio-grandi.
- ❑ Approccio consigliato da Inmon.



La Scelta dell'Architettura

- ❑ Solo dati pre-aggregati o anche dati di dettaglio?
- ❑ Quanti Modelli dei Dati?
 - ❑ Un solo Modello: il (multi)-dimensionale
 - ❑ Due Modelli distinti
 - E/R per i dati di dettaglio
 - (Multi-)dimensionale per i dati aggregati

- ❑ La scelta dell'Architettura
 - ❑ Influisce sulle scelte metodologiche
 - ❑ Ma non deve impedire flessibilità
 - Tutti gli approcci oggi consigliati prevedono una realizzazione incrementale
 - E' questione di ... metodo...
- ❑ Il Metodo deve mediare tra due estremi
 - ❑ Gli obiettivi di business espressi dagli utenti
 - ❑ L'effettiva disponibilità dei dati a supporto
- ❑ Ignorare gli uni o gli altri conduce facilmente al fallimento...

Una Questione di Metodo...

- ❑ Qualunque realizzazione incrementale offre il fianco al rischio di carenze di integrazione
- ❑ Il rischio si può minimizzare se
 - ❑ La sponsorizzazione del progetto è autorevole
 - ❑ C'è un forte controllo di progetto
 - Si eviti l'anarchia
 - Si eviti un eccessivo livello di parallelizzazione
 - ❑ C'è condivisione di metodo
 - ❑ C'è chiarezza negli obiettivi e nelle priorità

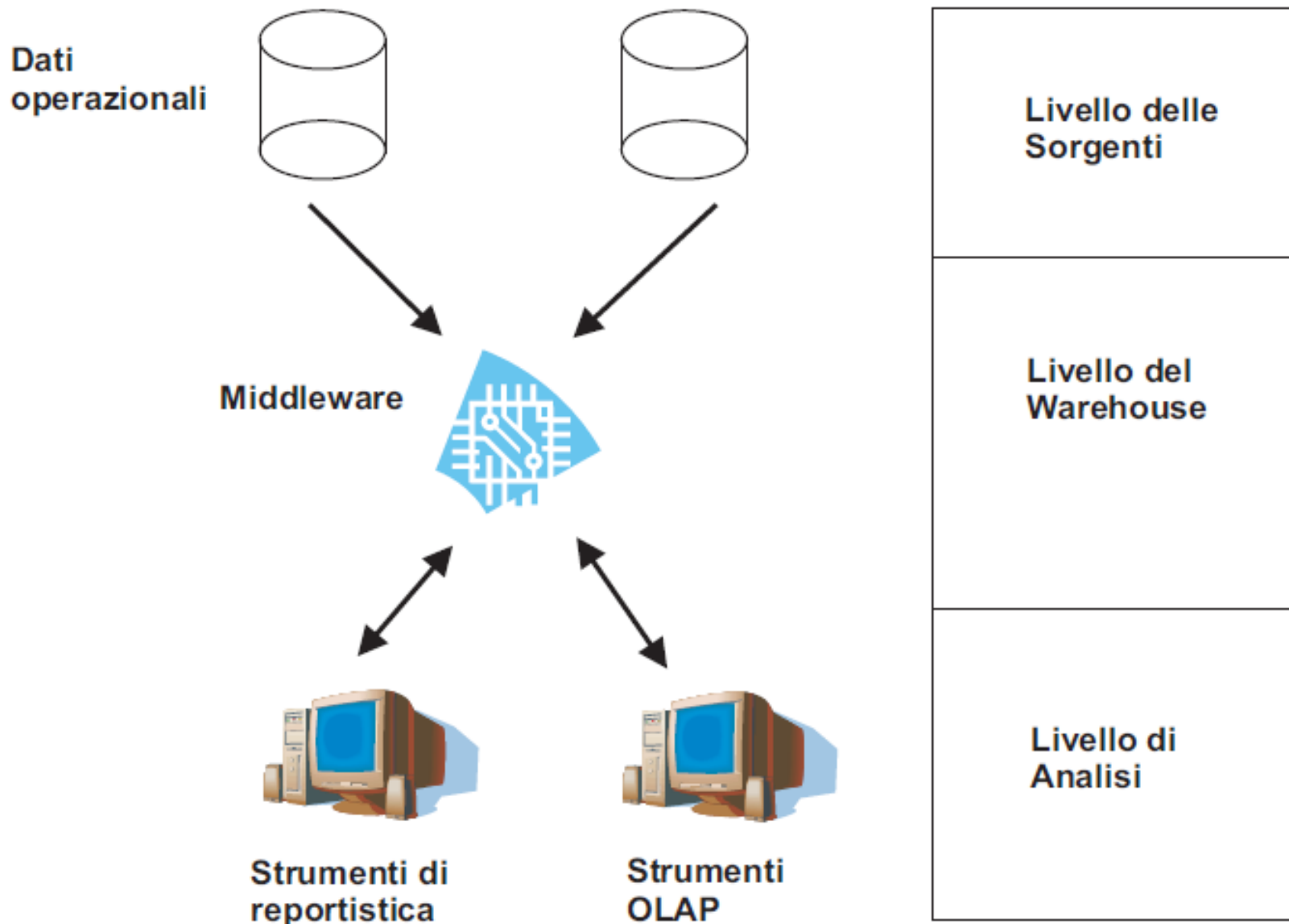
MODULE 11

ARCHITETTURE

- ❑ Le caratteristiche architettoniche irrinunciabili per un sistema di Data Warehousing possono essere così enunciate:
 - ❑ Separazione: l'elaborazione analitica e quella transazionale devono essere mantenute il più possibile separate.
 - ❑ Scalabilità: l'architettura hardware e software deve poter essere facilmente ridimensionata a fronte della crescita nel tempo dei volumi di dati da gestire ed elaborare e del numero di utenti da soddisfare.
 - ❑ Estendibilità: deve essere possibile accogliere nuove applicazioni e tecnologie senza riprogettare integralmente il sistema.
 - ❑ Sicurezza: il controllo sugli accessi è essenziale a causa della natura strategica dei dati memorizzati.
 - ❑ Amministrabilità: la complessità dell'attività di amministrazione non deve risultare eccessiva.

- ❑ Le caratteristiche architettoniche irrinunciabili per un sistema di Data Warehousing possono essere così enunciate:
 - ❑ Separazione: l'elaborazione analitica e quella transazionale devono essere mantenute il più possibile separate.
 - ❑ Scalabilità: l'architettura hardware e software deve poter essere facilmente ridimensionata a fronte della crescita nel tempo dei volumi di dati da gestire ed elaborare e del numero di utenti da soddisfare.
 - ❑ Estendibilità: deve essere possibile accogliere nuove applicazioni e tecnologie senza riprogettare integralmente il sistema.
 - ❑ Sicurezza: il controllo sugli accessi è essenziale a causa della natura strategica dei dati memorizzati.
 - ❑ Amministrabilità: la complessità dell'attività di amministrazione non deve risultare eccessiva.

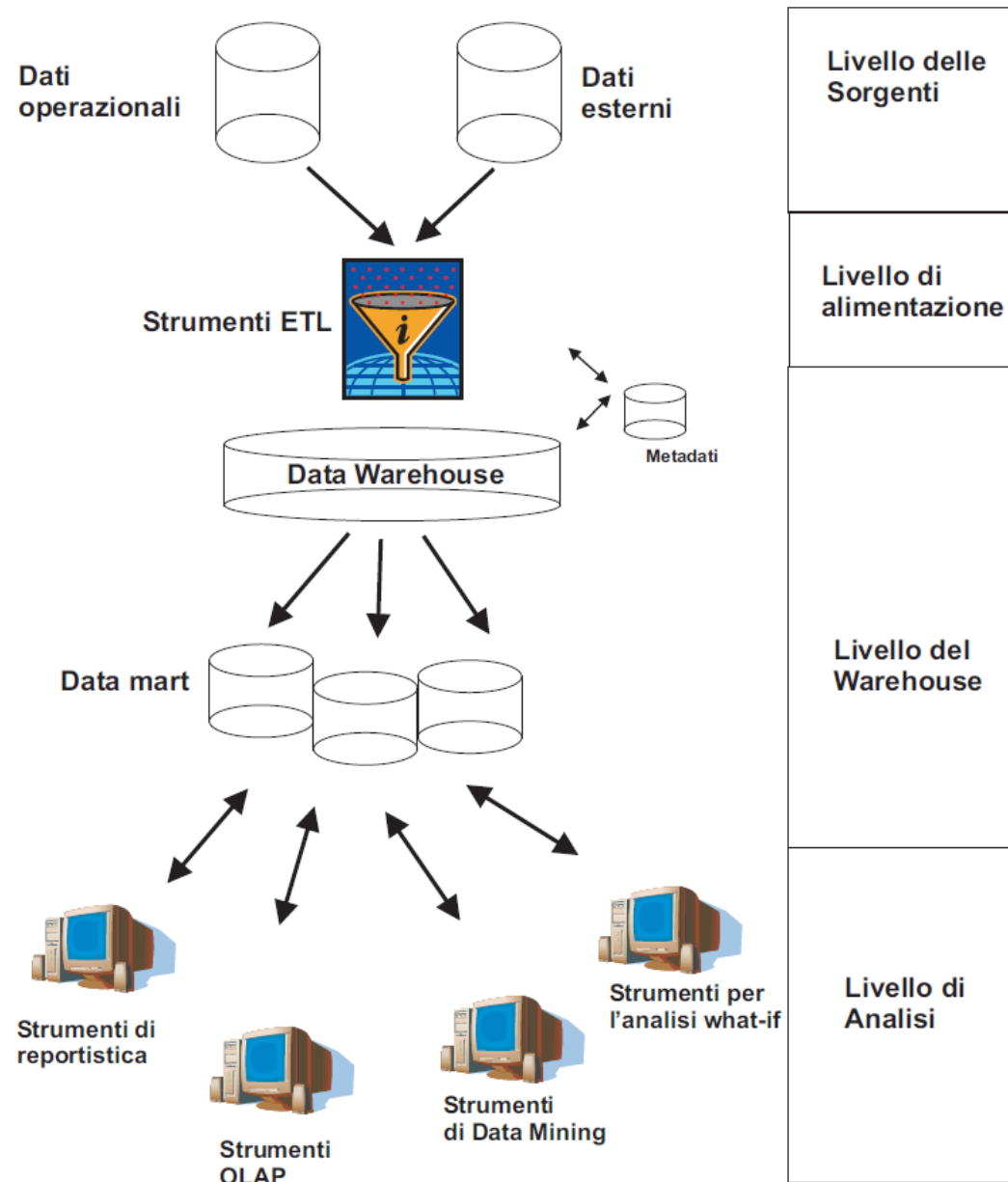
Architettura ad un livello – Grafico



Architettura ad un livello

- ❑ Obiettivo di questa architettura, a dire il vero poco utilizzata nella pratica, è la minimizzazione dei dati memorizzati, ottenuta eliminando le ridondanze.
- ❑ Il Data Warehouse è in questo caso virtuale, nel senso che viene implementato come una vista multidimensionale dei dati operazionali generata da un apposito middleware, ovvero da uno strato di elaborazione intermedio.
- ❑ Il primo punto debole di questa architettura è che non rispetta il requisito di separazione tra l'elaborazione analitica OLAP e quella transazionale OLTP. Le interrogazioni di analisi vengono, infatti, redirette sui dati operazionali dopo essere state reinterpretate dal middleware, interferendo, così, con il normale carico di lavoro transazionale.
- ❑ Inoltre, mentre con questa architettura è possibile (anche se complesso) rispondere ai requisiti di integrazione e correttezza dei dati, diventa impossibile esprimere un livello di storicizzazione superiore a quello delle sorgenti.
- ❑ Per questi motivi, l'approccio virtuale al Data Warehousing ha avuto successo soltanto in contesti in cui le esigenze di analisi sono particolarmente limitate e il volume dei dati da esaminare è molto ampio.

Architettura a due livelli – Grafico



Architettura a due livelli

- ❑ Il requisito di separazione gioca un ruolo fondamentale nel determinare la classica architettura di un sistema di DW
- ❑ Sebbene tradizionalmente denominata architettura a due livelli, per evidenziarne la separazione tra il livello delle sorgenti e quello del DW, essa, in realtà, si articola complessivamente su quattro livelli distinti, che descrivono stadi successivi del flusso di dati.
- ❑ Tali livelli sono:
 - ❑ Livello delle sorgenti. Il DW utilizza fonti di dati eterogenei estratti dall'ambiente di produzione e, quindi, archiviati originariamente in database aziendali, relazionali o legacy, oppure provenienti da sistemi informativi esterni all'azienda.
 - ❑ Livello dell'alimentazione. I dati memorizzati nelle sorgenti devono essere estratti, ripuliti (per eliminare le inconsistenze e completare eventuali parti mancanti) e integrati (per fondere sorgenti eterogenee secondo uno schema comune). I cosiddetti strumenti ETL (Extraction, Transformation and Loading) permettono di integrare schemi eterogenei, nonch'è di estrarre, trasformare, ripulire, validare, filtrare e caricare i dati dalle sorgenti nel DW. Dal punto di vista tecnologico vengono trattate problematiche tipiche dei servizi informativi distribuiti, come la gestione di dati inconsistenti

❑ e delle strutture dati incompatibili

- ❑ Il DW utilizza fonti di dati eterogenei estratti dall'ambiente di produzione e, quindi, archiviati originariamente in database aziendali, relazionali o legacy, oppure provenienti da sistemi informativi esterni all'azienda.

Architettura a due livelli – Livello dell'alimentazione

- ❑ I dati memorizzati nelle sorgenti devono essere estratti, ripuliti (per eliminare le inconsistenze e completare eventuali parti mancanti) e integrati (per fondere sorgenti eterogenee secondo uno schema comune).
- ❑ I cosiddetti strumenti ETL (Extraction, Transformation and Loading) permettono di integrare schemi eterogenei, nonch e di estrarre, trasformare, ripulire, validare, filtrare e caricare i dati dalle sorgenti nel DW.
- ❑ Dal punto di vista tecnologico vengono trattate problematiche tipiche dei servizi informativi distribuiti, come la gestione di dati inconsistenti e delle strutture dati incompatibili.

Architettura a due livelli – Livello del Warehouse

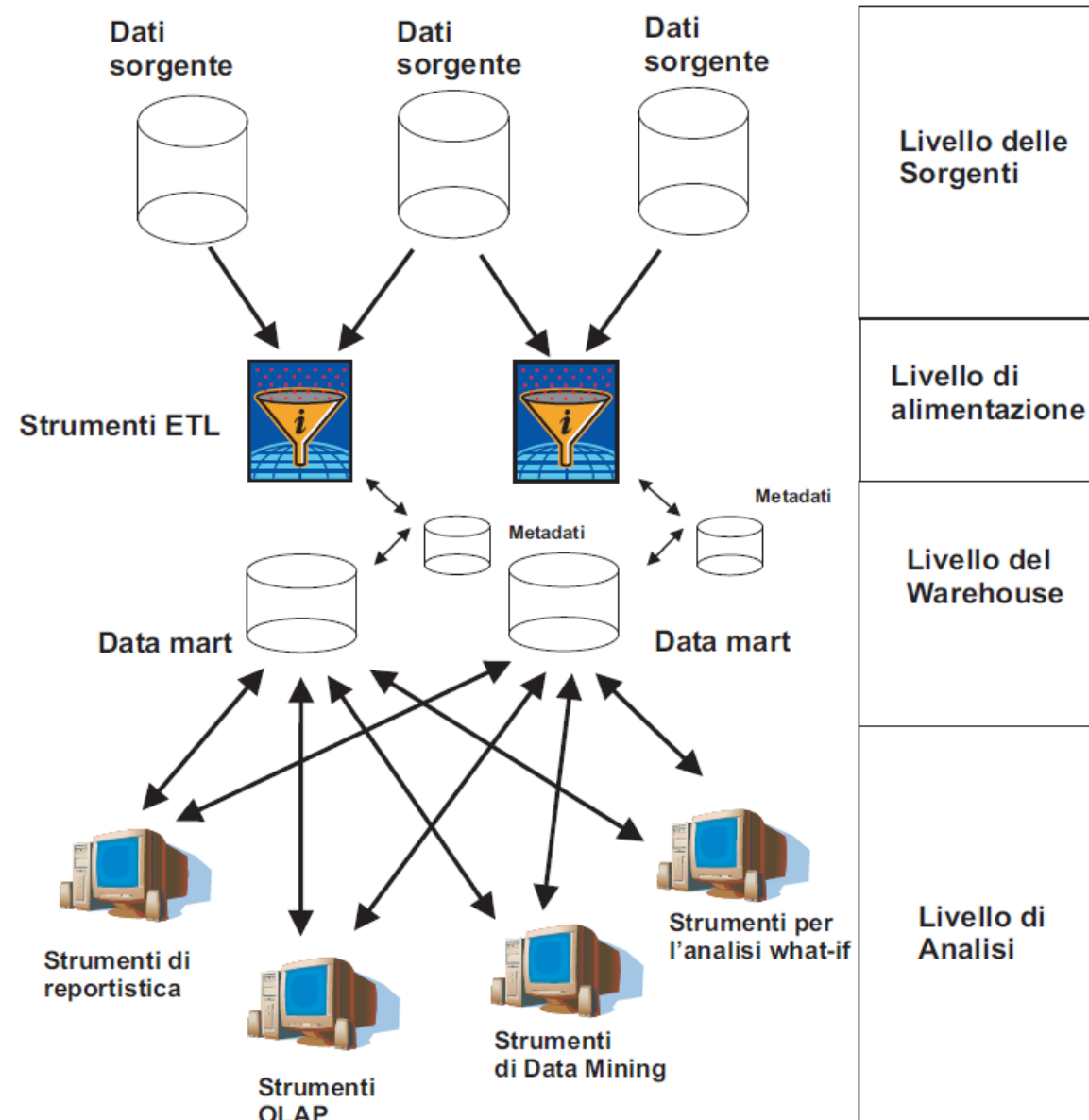
- ❑ Le informazioni vengono raccolte in un singolo “contenitore” (il Data Ware-house), centralizzato logicamente.
- ❑ Esso può essere direttamente consultato, ma anche usato come sorgente per costruire data mart; questi ultimi sono orientati verso specifiche aree dell’impresa e, di fatto, costituiscono una replica parziale del Data Warehouse.
- ❑ Accanto al Data Warehouse, il contenitore dei metadati mantiene informazioni sulle sorgenti, sui meccanismi di accesso, sulle procedure di pulizia ed alimentazione, sugli utenti, sugli schemi dei data mart, ecc.

- ❑ Permette la consultazione efficiente e flessibile dei dati integrati per la stesura dei report nonch'è per le attività di analisi e di simulazione.

- ❑ Dal punto di vista tecnologico sono richieste capacità di gestione dei dati aggregati, ottimizzazione di interrogazioni complesse, tecniche di indicizzazione avanzate e interfacce visuali amichevoli.
- ❑ La distinzione tra Data Warehouse e Data Mart, introdotta nell'architettura, merita un approfondimento.
- ❑ Il blocco che va sotto il nome di Data Warehouse viene, spesso, denominato anche Data Warehouse primario o Data Warehouse aziendale, e svolge il ruolo di contenitore centrale e globale dei dati di sintesi.
- ❑ I Data Mart possono essere visti come piccoli DW locali che replicano (ed eventualmente sintetizzano ulteriormente) la porzione di DW primario di interesse per una particolare area applicativa.

- ❑ Con il termine Data Mart si intende un sottoinsieme o un'aggregazione dei dati presenti nel DW primario, contenente l'insieme delle informazioni rilevanti per una particolare area del business, una particolare divisione dell'azienda, oppure una particolare categoria di soggetti.
- ❑ I Data Mart alimentati dal DW primario sono spesso detti dipendenti. Sebbene, in linea di principio non strettamente necessari, per i sistemi collocati all'interno di realtà aziendali medio-grandi essi costituiscono un'utilissima risorsa:
 - ❑ come blocchi costruttivi durante la realizzazione incrementale del DW;
 - ❑ in quanto delineano i contorni delle informazioni necessarie ad una particolare tipologia di utenti;
 - ❑ in quanto permettono di raggiungere prestazioni migliori, essendo di dimensioni inferiori al DW primario.
- ❑ In alcuni contesti, principalmente per motivi organizzativi e politici, si preferisce adottare un'architettura in cui i data mart vengono alimentati direttamente dalle sorgenti e vengono, pertanto, detti indipendenti.
- ❑ L'assenza di un DW primario snellisce le fasi progettuali ma determina uno schema complesso per l'accesso ai dati e ingenera il rischio di inconsistenza tra i Data Mart.

Architettura a due livelli con Data Mart indipendenti

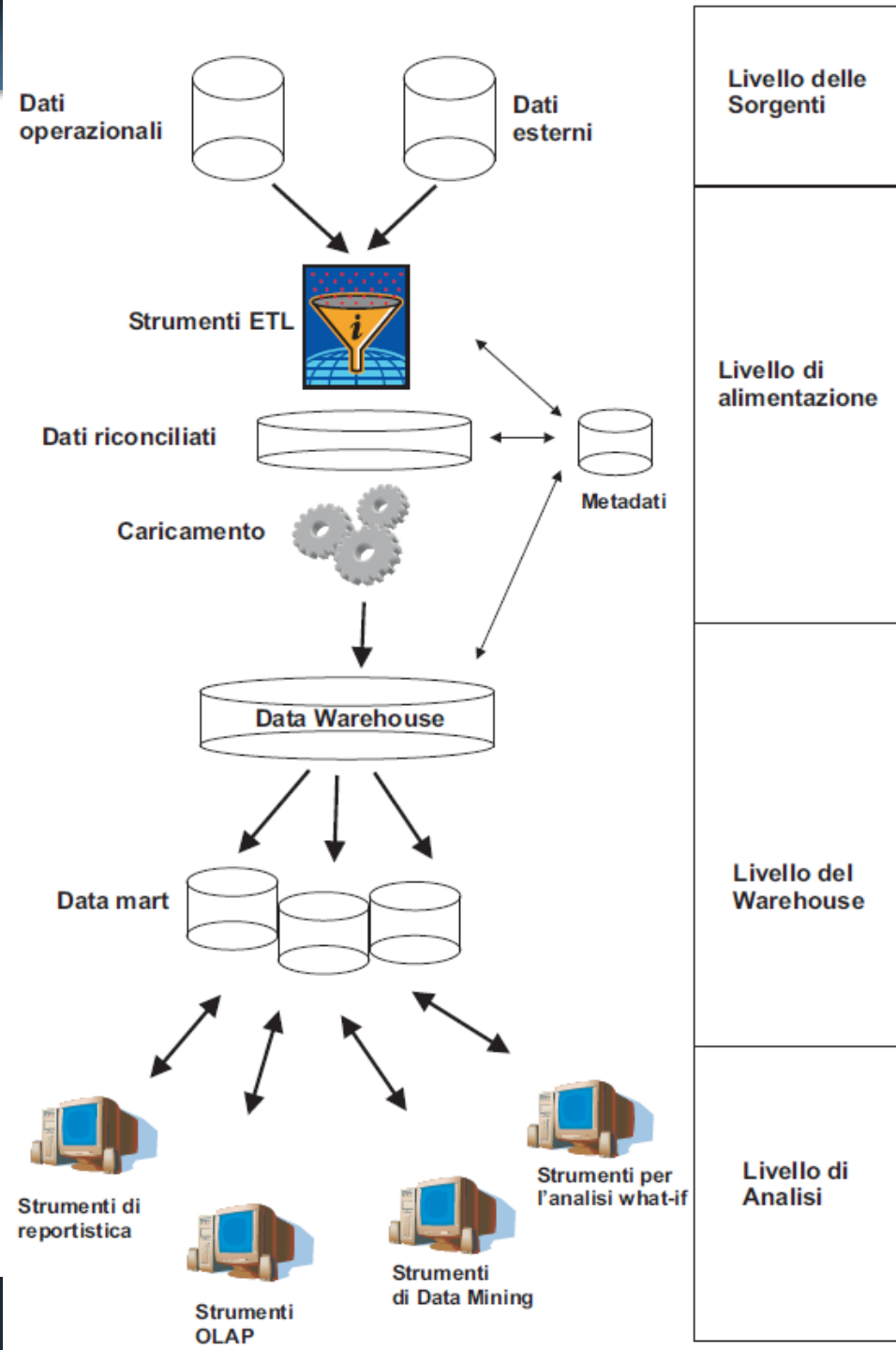


- ❑ Per tale ragione, a volte, pur rispettando l'indipendenza dei Data Mart, si preferisce creare comunque un DW centrale; in quest'ultimo caso, rispetto all'architettura standard a due livelli, i ruoli dei Data Mart e del DW sono, di fatto, invertiti; infatti, il DW, in questo caso, viene alimentato dai Data Mart e può essere direttamente interrogato al fine di semplificare l'accesso ai dati.

- ❑ Le principali motivazioni a sostegno dell'architettura a due livelli, in cui il livello del warehouse funge da separatore tra le sorgenti e le applicazioni di analisi, sono così riassumibili:
 - ❑ A livello del DW è continuamente disponibile informazione di buona qualità anche quando, per motivi tecnici oppure organizzativi, è temporaneamente precluso l'accesso alle sorgenti.
 - ❑ L'interrogazione analitica effettuata sul DW non interferisce con la gestione delle transazioni a livello operativo, la cui affidabilità è essenziale per il funzionamento dell'azienda.
 - ❑ L'organizzazione logica del DW è basata sul modello multidimensionale, mentre le sorgenti presentano, in genere, modelli relazionali o semi-strutturati.
 - ❑ C'è una discordanza temporale e di granularità tra sistemi OLTP, che trattano dati correnti e al massimo livello di dettaglio, e sistemi OLAP, che operano su dati storici e di sintesi.
 - ❑ A livello del DW è possibile utilizzare tecniche specifiche per ottimizzare le prestazioni per applicazioni di analisi e reportistica.

- ❑ E' utile osservare che alcuna, con riferimento all'architettura appena analizzata, utilizzano la stessa terminologia per indicare concetti differenti.
- ❑ In particolare, essi considerano il DW come un contenitore di dati integrati e consistenti, ma ancora in forma operativa, introducendo la rappresentazione multidimensionale dei dati solo a livello dei Data Mart.
- ❑ Nella nostra terminologia questa visione operativa del DW corrisponde, sostanzialmente, al livello dei dati riconciliati nelle architetture a tre livelli, trattate prossime slides.

Architettura a tre livelli



- ❑ Il terzo livello introdotto in questa architettura è il cosiddetto livello dei dati riconciliati, detto anche operational data store, che materializza i dati operazionali ottenuti a valle del processo di integrazione e ripulitura dei dati sorgente: si ottengono, quindi dati integrati, consistenti, corretti, volatili, correnti e dettagliati.
- ❑ Come illustrato nella slide precedente, il DW non viene più alimentato direttamente dalle sorgenti, bensì dai dati riconciliati.
- ❑ Il vantaggio principale del livello dei dati riconciliati `e che esso crea un modello di dati comune e di riferimento per l'intera azienda, introducendo, al contempo, una separazione netta tra le problematiche legate all'estrazione e integrazione dei dati dalle sorgenti e quelle inerenti l'alimentazione del DW.
- ❑ D'altro canto, i dati riconciliati introducono un'ulteriore ridondanza rispetto ai dati operazionali sorgente.
- ❑ Va, comunque, detto che, in realtà, si può assumere che anche nelle architetture a due livelli sia presente un livello riconciliato, che non sarà, in quel caso, materializzato ma soltanto virtuale, essendo definito come una vista integrata e consistente dei dati memorizzati nelle sorgenti operazionali.

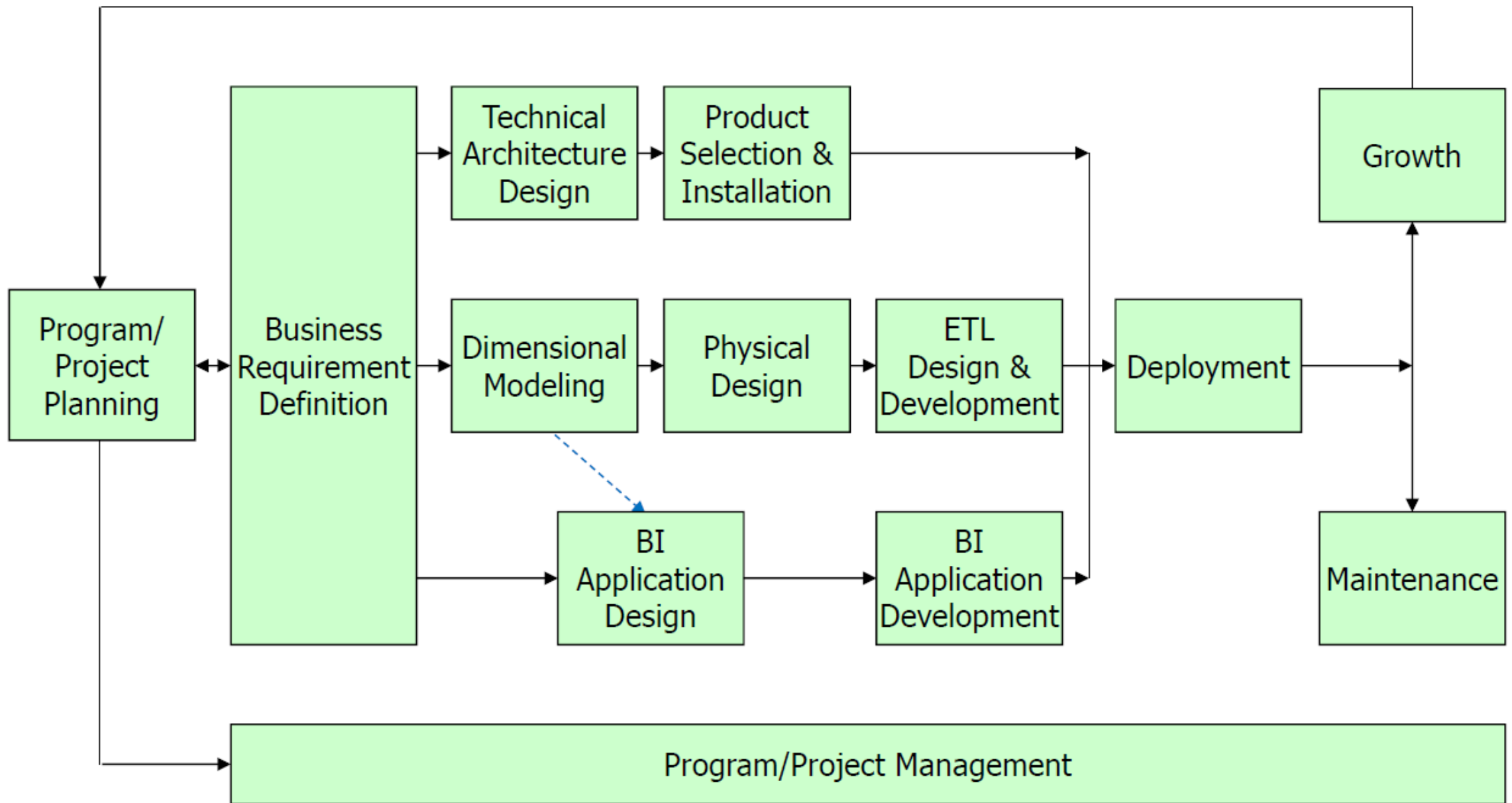
MODULO 12

CICLO DI VITA DIMENSIONALE

- ❑ Il ciclo di vita dimensionale – Business Dimensional Lifecycle, chiamato anche Kimball Lifecycle – descrive il framework complessivo che lega le diverse attività dello sviluppo di un sistema DW/BI
- ❑ è una metodologia completa per la progettazione e realizzazione di un sistema DW
- ❑ fornisce il contesto di riferimento per la progettazione e realizzazione di data warehouse dimensionali

Caratteristiche del ciclo di vita dimensionale

- ❑ un sistema DW/BI viene costruito a partire dagli utenti di business e dai loro bisogni effettivi – con questi obiettivi in mente, si lavora “all’indietro” attraverso i report, le applicazioni, le basi di dati e il software – fino agli strati più fisici del sistema
- ❑ si tratta dunque di un approccio fortemente guidato dal business e dall’utente – e non dalle tecnologie (come avviene in altre metodologie)
- ❑ è un processo iterativo – in realtà, incrementale
- ❑ l’intero sistema DW/BI (“programma”) viene sviluppato in una serie di iterazioni/cicli/incrementi (“progetti”) – e non come un unico Big Bang galattico
- ❑ ciascun “progetto” ha lo scopo di fornire capacità di analisi (aggiuntive) di valore per gli utenti di business
- ❑ consistenza del sistema garantita mediante un’architettura a bus del data warehouse



Pianificazione del progetto e del programma

- ❑ Il ciclo di vita inizia con la pianificazione del programma e del progetto
 - ❑ per programma si intende l'attività complessiva di sviluppo di un sistema DW/BI completo, che procede tramite diverse iterazioni
 - ❑ per progetto si intende una singola iterazione/ ciclo/incremento del ciclo di vita, dall'avvio al rilascio
 - ❑ alcune attività della pianificazione
 - verifica di “readiness” dell'organizzazione
 - analisi preliminare dei requisiti e della portata del programma o del progetto
 - valutazioni economiche
 - identificazione delle risorse
 - formazione dei gruppi di lavoro
 - definizione del piano di sviluppo (iterativo)

- ❑ Per tutta la sua durata, il ciclo di vita richiede un'attività di gestione del progetto
 - ❑ le attività della gestione del progetto/programma si focalizzano sul monitoraggio dello stato del progetto, controllo dei cambiamenti
 - ❑ un aspetto critico nel raggiungimento degli obiettivi di un sistema DW/BI è la comunicazione continua
 - per questo, la gestione comprende anche lo sviluppo di un piano di comunicazione completo – che riguarda sia le parti interessate di business che quelle IT

- ❑ Un sistema DW/BI viene costruito a partire dagli utenti di business e dai loro bisogni effettivi
 - ❑ una definizione efficace dei requisiti di business è cruciale –poiché stabilisce le fondamenta per tutte le attività del ciclo di vita successive
 - ❑ pertanto, è fondamentale una comprensione accurata delle necessità dei diversi utenti di business e dei loro requisiti
 - ❑ due piani principali
 - comprendere i bisogni e le priorità degli utenti a livello dell'intero programma di sviluppo – probabilmente, a livello dell'intera organizzazione
 - comprendere in dettaglio i requisiti di business nel contesto di un singolo progetto – la cui portata può essere abbastanza ristretta
- ❑ importante coinvolgere anche le parti interessate dell'IT

- ☐ Alcune domande per comprendere i requisiti di business al livello dell'intero programma
 - ☐ quali gli obiettivi della tua organizzazione? quali sono i processi di business fondamentali della tua organizzazione?
 - ☐ come misuri il successo? come fai a capire se stai raggiungendo i tuoi obiettivi? quanto spesso ti misuri?
 - ☐ quali sono le questioni chiave del tuo business oggi? Come le affronti? che impatto hanno sulla tua organizzazione?
 - ☐ come identifichi i problemi o capisci che ci potrebbero essere quali opportunità ci sono nel tuo business, che potrebbero avere un impatto importante per la tua organizzazione? Con quale impatto economico?
 - ☐ qual è la tua visione sul come sfruttare al meglio le informazioni nella tua organizzazione?

Raccolta e analisi dei requisiti (progetto)

- ❑ Ciascuna iterazione/progetto del ciclo di vita si focalizzerà invece su un singolo processo di business (o iniziativa strategica di business) dell'organizzazione – alcune domande importanti
 - ❑ quali sono i prodotti di questo processo? come si possono distinguere i diversi prodotti? quanto spesso cambiano?
 - ❑ quali analisi vengono fatte per valutare questo processo? con quali dati? quale uso fai di questi dati quando ne vieni in possesso? quali analisi vorresti invece poter eseguire?
 - ❑ che report usi/prepari? quali i dati più importanti dei report? quali report vorresti invece? quali report non guardi mai?
 - ❑ quanta informazione storica è necessaria? quanto spesso deve essere aggiornata? per quanto tempo deve essere mantenuta?
 - ❑ quali opportunità esistono per migliorare in modo significativo il tuo business sulla base di un accesso migliore alle informazioni? con quale impatto economico?

- ❑ La progettazione – nonché il relativo sviluppo – del sistema DW/DI avviene su tre tracce concorrenti
 - ❑ progettazione dei dati
 - progettazione dimensionale (logica) e fisica del data warehouse
 - progettazione e sviluppo del sottosistema ETL
 - ❑ progettazione tecnologica
 - progettazione dell'architettura tecnica del data warehouse
 - selezione dei prodotti e loro installazione
 - ❑ progettazione delle applicazioni BI
 - specifica delle applicazioni BI
 - sviluppo delle applicazioni BI

- ❑ La progettazione dimensionale riguarda la progettazione logica dei dati del data warehouse – costituisce la pietra angolare della progettazione dell'intero sistema DW/BI
- ❑ i dati per l'analisi di ciascun processo di business di interesse nel progetto corrente vengono rappresentati mediante un modello dimensionale (o data mart) per quel processo di business
- ❑ la matrice a bus del data warehouse rappresenta tutti i processi di business fondamentali dell'organizzazione e le loro dimensioni principali di analisi – questa matrice viene definita inizialmente durante l'analisi iniziale dei requisiti di business – e poi raffinata in modo iterativo
 - questa matrice serve a garantire che i dati nei modelli dimensionali progettati nelle diverse iterazioni/progetti possano essere correlati e integrati nel corso del tempo e dell'intero programma

- ❑ Definizione delle strutture fisiche adeguate a una rappresentazione efficiente dello schema logico del data warehouse
 - ❑ definizione dello schema fisico dei dati – sia a livello relazionale che OLAP
 - ❑ progettazione delle strategie di tuning delle prestazioni – dall'indicizzazione al partizionamento dei dati alle aggregazioni

- ❑ Progetto e sviluppo del sottosistema ETL
 - ❑ 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
 - ❑ una delle attività più complesse con cui si deve confrontare il team di sviluppo
 - mediamente, il 70% dei rischi e dello sforzo di sviluppo in un progetto DW/BI sono relativi a questa attività

- ❑ Progetto e sviluppo del sottosistema ETL
 - ❑ lavorazione delle dimensioni
 - estrazione dei dati delle dimensioni dalle loro sorgenti informative
 - pulizia dei valori degli attributi – parsing, decodifica, gestione di dati mancanti o non validi
 - gestione dei cambiamenti delle dimensioni
 - assegnazione delle chiavi surrogate
 - caricamento delle nuove dimensioni – e pubblicazione della loro avvenuta revisione

❑ Progetto e sviluppo del sottosistema ETL

❑ lavorazione dei fatti

- estrazione dei dati sui fatti dalle loro sorgenti informative
- ricezione delle dimensioni aggiornate
- separazione dei fatti in base alla loro granularità
- trasformazione dei fatti, come richiesto
- sostituzione delle chiavi nelle sorgenti informative con le chiavi surrogate delle dimensioni
- aggiunta di nuove chiavi per dimensioni supplementari – che descrivono il contesto dei fatti
- verifica di qualità della tabella fatti
- costruzione o aggiornamento di tabelle fatti aggregate
- caricamento dei dati nelle tabelle fatti – e pubblicazione del loro avvenuto aggiornamento

- ❑ Progettazione dell'architettura tecnica globale del sistema DW/BI
 - ❑ l'ambiente di un sistema DW/BI richiede l'integrazione di numerosi sistemi e tecnologie
 - i sistemi (legacy) su cui risiedono le sorgenti informative, su cui viene parzialmente eseguita la preparazione dei dati
 - i sistemi (solitamente aperti e basati su tecnologia relazionale) su cui risiedono i server di presentazione
 - i sistemi client
 - ❑ l'architettura tecnica è influenzata da tre fattori
 - requisiti di business – l'ambiente tecnico attuale – le direzioni strategie pianificate per l'ambiente tecnico
- ❑ Selezione dei prodotti e loro installazione
 - ❑ scelta degli ambienti hardware e software

- ❑ La “front room” del sistema DW/BI è composta da un insieme di applicazioni di business intelligence (BI) – utilizzate direttamente dagli utenti di business del DW/BI
- ❑ gli utenti del sistema possono avere necessità di analisi molto diverse tra loro
 - per questo, sono necessari diversi tipi di applicazioni BI – che variano da semplici report standard parametrizzati a sistemi complessi di analisi dei dati
- ❑ in generale, le applicazioni BI supportano un processo ciclico comune nelle analisi di business
 - dal monitoraggio dei processi di business all’identificazione di un problema o di un’opportunità, dal determinare l’azione da intraprendere al monitorare il risultato di quell’azione

- ❑ Le tre tracce concorrenti di progettazione (dei dati, tecnologica e delle applicazioni BI) convergono nella fase di deployment
 - ❑ installazione del sistema
 - ❑ integrazione dei sottosistemi e testing
 - ❑ formazione degli utenti
 - ❑ documentazione e supporto agli utenti

- ❑ La manutenzione comprende
 - ❑ monitoraggio delle qualità del sistema
 - ❑ misurazione delle prestazioni e dell'uso
 - ❑ ottimizzazione tecnica
 - ❑ tuning delle prestazioni, manutenzione degli indici, backup di sistema
- ❑ Un sistema DW/BI realizzato con il ciclo di vita dimensionale, se ben fatto, è pronto a evolvere e a crescere – con nuovi progetti, per fornire sempre più valore ai suoi utenti
- ❑ l'evoluzione di un data warehouse è indice di successo, non di fallimento
- ❑ ciascun progetto riguarda di solito il miglioramento dei modelli dimensionali esistenti, ma anche e soprattutto lo sviluppo di modelli dimensionali per ulteriori processi di business