

# Introduzione alla progettazione di data warehouse dimensionali

Luca Cabibbo  
marzo 2010

## Progettazione di data warehouse dimensionali

Questo corso è un'introduzione alla progettazione di data warehouse dimensionali

- contenuti del corso
  - modello dimensionale e data warehouse dimensionali
  - tecniche di modellazione dimensionale
  - aggregazioni
  - ciclo di vita dimensionale dei data warehouse
  - preparazione dei dati
  - applicazioni per il data warehouse

# 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

- Data Warehouse – La guida completa

- Ralph Kimball & Margy Ross
- Hoepli, 2002

attenzione, purtroppo la traduzione è pessima

- The Data Warehouse Lifecycle Toolkit

- Ralph Kimball et al.
- John Wiley & Sons, 1998

Questo materiale è disponibile online al sito

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

# Che cosa è un data warehouse

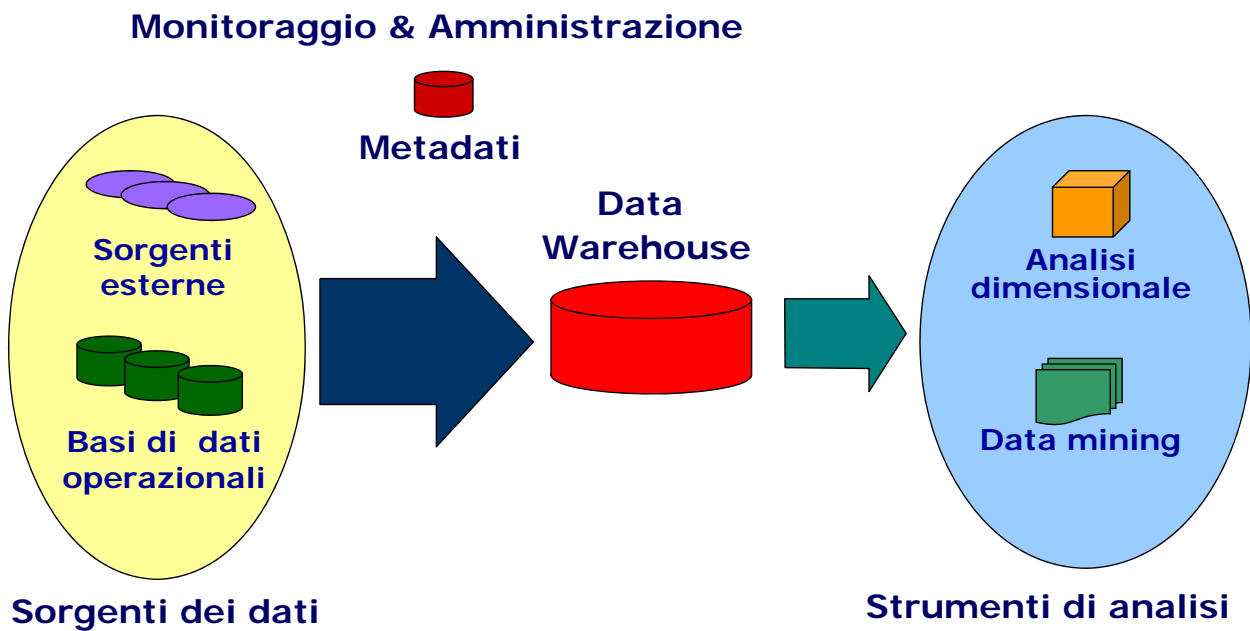
Un **data warehouse** è una base di dati

- 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

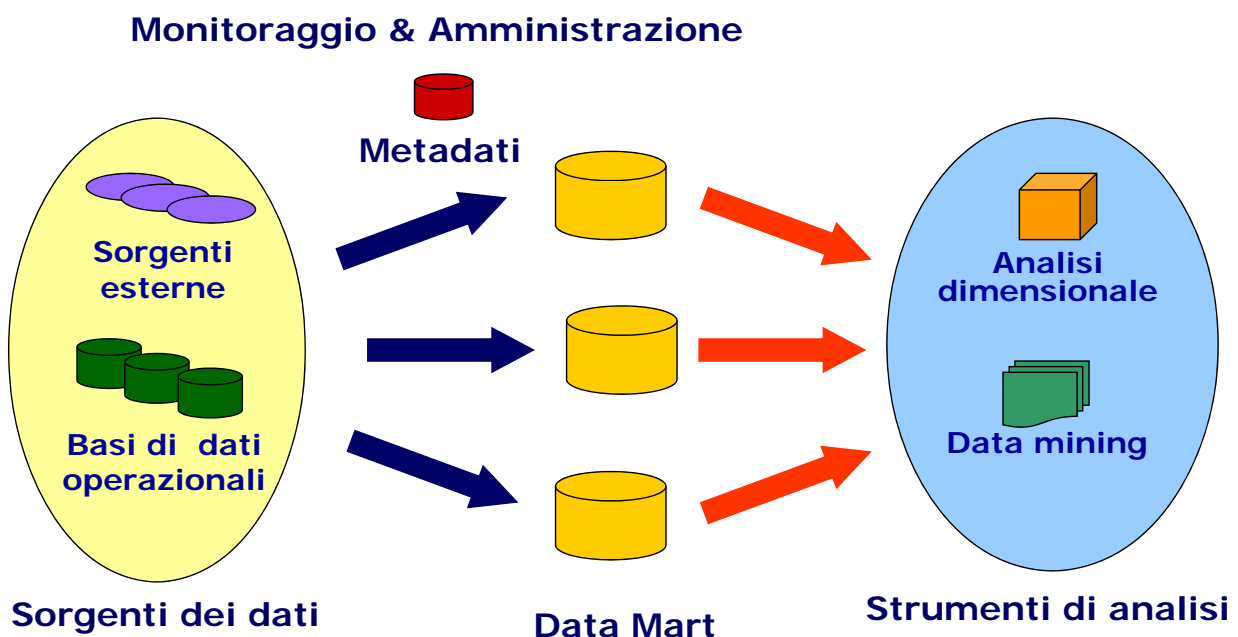


5

Introduzione alla progettazione di data warehouse dimensionali

Luca Cabibbo

# Architettura per il data warehousing (Kimball)

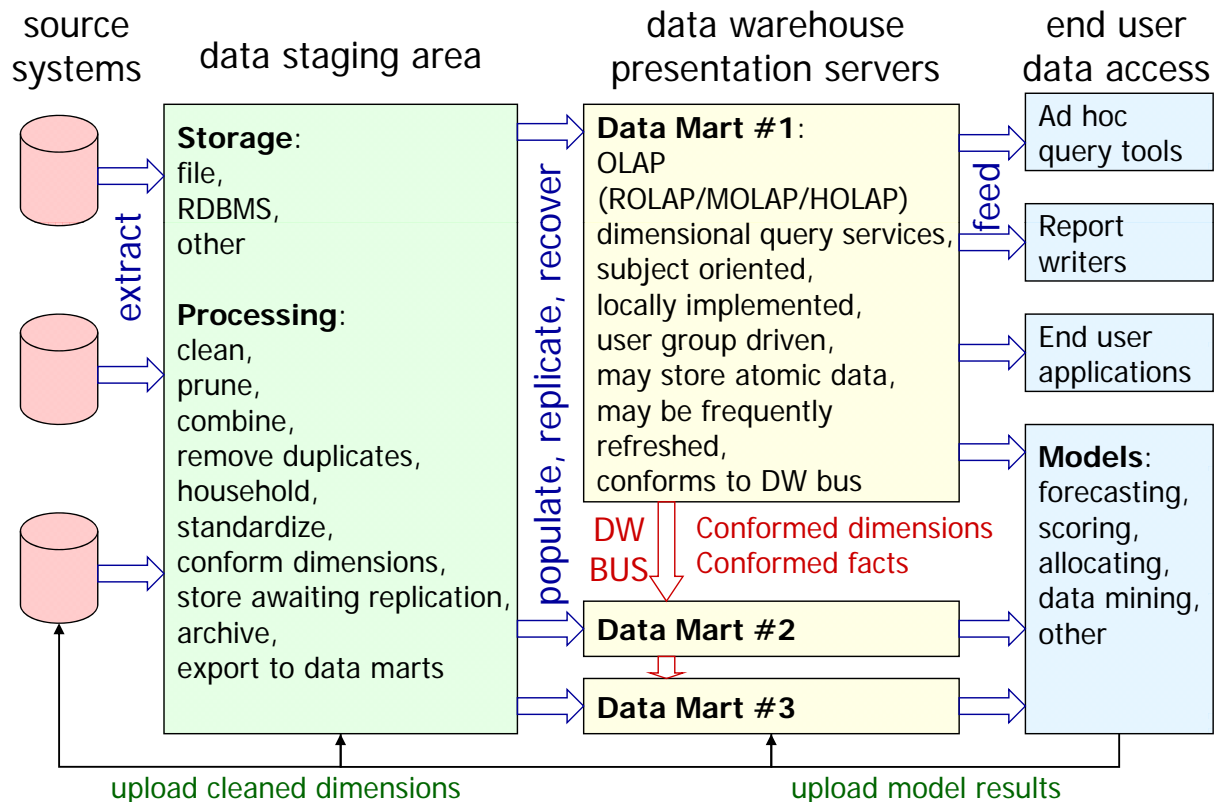


6

Introduzione alla progettazione di data warehouse dimensionali

Luca Cabibbo

# Elementi di un data warehouse



7

Introduzione alla progettazione di data warehouse dimensionali

Luca Cabibbo

## Sorgenti informative

Le **sorgenti informative** di un data warehouse comprendono

- i sistemi operazionali dell'organizzazione
  - sono sistemi transazionali (OLTP) orientati alla gestione dei processi operazionali
  - non mantengono dati storici
  - ogni sistema gestisce uno o più "soggetti" (ad esempio, prodotti o clienti)
    - nell'ambito di un processo
    - normalmente non in modo standardizzato (conforme) nell'ambito dell'organizzazione
  - possono essere sistemi "legacy"
- sorgenti esterne
  - ad esempio, dati forniti da società specializzate di analisi

8

Introduzione alla progettazione di data warehouse dimensionali

Luca Cabibbo

## Area di preparazione dei dati

L'area di **preparazione dei dati (data staging)** è usata per il transito dei dati dalle sorgenti informative al data warehouse

- comprende ogni cosa tra le sorgenti informative e i server di presentazione
  - aree di memorizzazione dei dati estratti dalle sorgenti informative e preparati per il caricamento nel data warehouse
  - processi per la preparazione di tali dati
    - pulizia, trasformazione, combinazione, rimozione di duplicati, archiviazione, preparazione per l'uso nel data warehouse
- la preparazione dei dati è un insieme complesso di attività semplici
- è distribuita su più calcolatori e ambienti eterogenei
- i dati sono memorizzati prevalentemente su file

## Server di presentazione

Un **server di presentazione** è un sistema in cui i dati del data warehouse sono organizzati e memorizzati per essere interrogati direttamente da utenti finali, report writer e altre applicazioni

- i dati sono rappresentati in forma dimensionale
  - ovvero, sono organizzati usando i concetti di fatto e dimensione
- tecnologie che possono essere adottate
  - RDBMS
    - i dati sono organizzati mediante degli schemi dimensionali (schemi a stella)
  - tecnologia OLAP
    - i concetti di fatto e dimensione sono espliciti
  - tecnologie ibride – RDBMS con estensioni OLAP

## Data warehouse

Il **data warehouse** è la base di dati dell'organizzazione che viene usata per il supporto alle decisioni

- è una sorgente di dati interrogabile
  - è organizzata e ottimizzata per supportare elaborazioni di tipo analitico (OLAP)
- è aggiornata frequentemente e in modo controllato
  - i dati provengono dall'area di preparazione dei dati
  - viene aggiornata quando
    - vengono accumulati degli snapshot dei dati
    - i dati vengono corretti
- supporta quindi due modalità di funzionamento

## Accesso ai dati

Gli utenti finali del data warehouse accedono ai dati del data warehouse usando una varietà di strumenti

- report writer
  - uno strumento per la generazione di rapporti standard
- applicazioni per l'utente finale finale (end user application)
  - un insieme di strumenti per interrogare, analizzare e presentare dati in un specifico processo
- ad hoc query tool
  - l'utente esprime le proprie interrogazioni manipolando (più o meno direttamente) gli schemi dimensionali
- applicazioni di modellazione
  - un client sofisticato con capacità analitiche

## OLAP

**OLAP** (On Line Analytical Processing) è l'attività di interrogazione e presentazione dei dati di un data warehouse in uno stile dimensionale

Il termine OLAP indica anche una tecnologia (lato server e lato client) che consente una rappresentazione dei dati basata su fatti e dimensioni

- per MOLAP (Multidimensional OLAP) si intende la tecnologia multidimensionale, in cui i dati sono fisicamente rappresentati sotto forma di cubo multidimensionale
- per ROLAP (Relational OLAP) si intende l'estensione della tecnologia relazionale con i concetti dimensionali

Qualunque sia la tecnologia usata, in qualche fase i dati devono essere rappresentati in modo conforme al modello dimensionale

## Metadati

I **metadati** sono tutte le informazioni del sistema di data warehousing che non corrispondono a dati memorizzati dal data warehouse

- molte delle attività in un data warehouse devono essere parametrizzate e misurate
  - i valori effettivi dei parametri e le misure relative all'uso del data warehouse sono metadati che vanno opportunamente identificati, memorizzati, usati e controllati

## Sul data warehouse

Il **data warehouse** è la base di dati dell'organizzazione che viene usata per il supporto alle decisioni

- secondo Kimball, “il data warehouse non è nulla più che l'unione dei data mart che lo costituiscono”
- andiamo ad approfondire questo aspetto

## Processi di business

Un **processo di business** (o **processo aziendale**, **business process**) è un insieme coerente di attività significative per l'utente del data warehouse

- definizione volutamente vaga – data la generalità della nozione di processo
- un processo è un insieme di attività
  - ad esempio, gestione ordini, inventario, consegna ai clienti, ...

Perché i processi sono importanti?

- i processi possono essere usati come criterio per il raggruppamento coerente delle risorse informative dell'organizzazione e dei dati del data warehouse
- un data warehouse viene realizzato mediante uno o più data mart per ciascun processo di interesse

## Data mart

Un **data mart** è un sottoinsieme logico dell'intero data warehouse

- un data mart è la restrizione del data warehouse a un singolo processo
- il data warehouse è l'unione dei data mart che lo costituiscono

Pro e contro dei data mart

- un data mart rappresenta un progetto solitamente fattibile
  - la realizzazione diretta di un data warehouse completo non è invece solitamente fattibile
- tuttavia, la realizzazione di un insieme di data mart non porta necessariamente alla realizzazione del data warehouse
  - necessaria una certa "coerenza" tra i diversi data mart

## Data warehouse dimensionali

Un **data warehouse dimensionale** è un data warehouse realizzato come unione di un insieme di data mart che hanno le seguenti caratteristiche

- ogni data mart è un insieme di schemi dimensionali
- viene adottata un'**architettura a bus del data warehouse** – **data warehouse bus architecture** – ovvero, i vari data mart sono costruiti usando
  - dimensioni conformi (o conformate)
    - ciascuna dimensione ha lo stesso significato in ciascuno schema dimensionale e data mart
  - fatti conformi
    - anche i fatti hanno un'interpretazione uniforme

## Dati multidimensionali

L'organizzazione logica dei dati in un data warehouse è descritta secondo un modello di dati (multi)dimensionale

- i dati sono descritti con riferimento ai concetti di fatto e dimensione
  - secondo la prospettiva degli utenti del data warehouse (analisti)
  - e non secondo i modelli (ad esempio, il modello relazionale) adottati nei sistemi che gestiscono le sorgenti informative

## Dati multidimensionali

L'analisi dei dati avviene su dati rappresentati in forma multidimensionale, ovvero organizzati mediante i seguenti concetti

- **fatto** (o processo)
  - un concetto sul quale centrare l'analisi
- **misura**
  - una proprietà atomica o misura di un fatto da analizzare
  - le misure sono solitamente valori numerici e additivi su un dominio continuo
- **dimensione**
  - una prospettiva rispetto alla quale effettuare l'analisi
  - le dimensioni descrivono domini discreti, solitamente organizzati in livelli di aggregazione

## Dati multidimensionali - esempi

### Data mart delle vendite

- fatto: vendite dei prodotti, giornaliera, per negozio
- dimensioni: prodotto, tempo (giorno), negozio, promozione
- misure: quantità venduta, incasso, costo, conteggio dei clienti

### Data mart delle telefonate

- fatto: telefonata
- dimensioni: chiamante, chiamato, tariffa, tempo (giorno), tempo (ora del giorno)
- misure: durata, costo

## Modello dimensionale

Nel corso viene adottato il **modello dimensionale**

- un modo di utilizzare una base di dati relazionale per rappresentare dati multidimensionali
- uno schema dimensionale è uno schema relazionale di forma particolare – star schema, o schema a stella
- lo schema di un data warehouse è un insieme di schemi dimensionali

## Schemi dimensionali

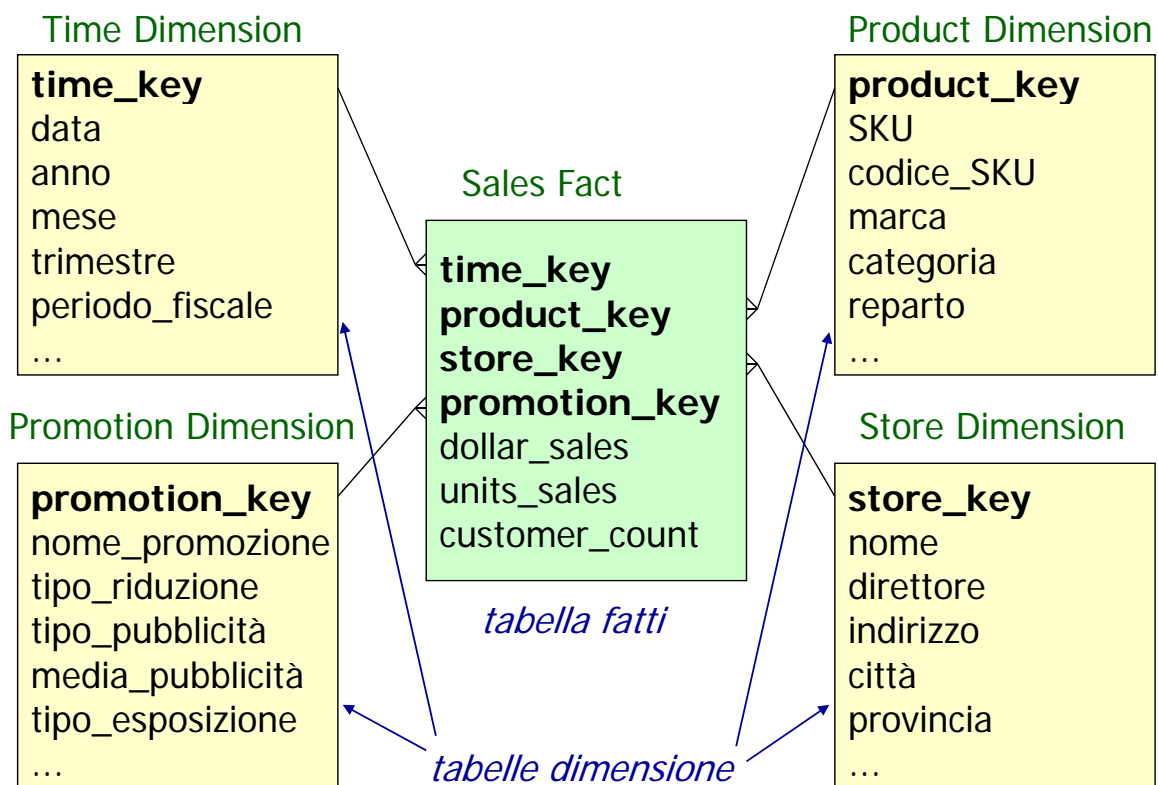
Uno **schema dimensionale** è composto da

- una tabella principale, chiamata **tabella fatti**
  - la tabella fatti memorizza i fatti misurabili di un processo
    - i fatti più comuni sono numerici, continui e additivi
- alcune tabelle ausiliarie, chiamate **tabelle dimensione**
  - una tabella dimensione rappresenta una dimensione rispetto alla quale è interessante analizzare i fatti
    - memorizza i membri che caratterizzano la grana dei fatti, nonché gli attributi (solitamente testuali, discreti e descrittivi) usati dalle interrogazioni per vincolare e raggruppare i fatti

ed ha lo scopo

- di rappresentare un insieme di funzioni
- sulle quali è di interesse eseguire operazioni di aggregazioni

## Uno schema dimensionale



# Attività in un data warehouse

Le attività di base in un data warehouse comprendono

- attività nell'area di preparazione dei dati – attività “notturne”
  - estrazione, trasformazione, caricamento e indicizzazione, controllo di qualità
  - aggiornamento del data warehouse
- attività utente – attività “diurne”
  - interrogazione
- attività di amministrazione
  - gestione della sicurezza
  - auditing
  - backup e recovery
  - gestione del feedback

## Estrazione

L'**estrazione** è il primo passo nel transito dei dati dalle sorgenti informative al data warehouse

- più precisamente, l'attività di estrazione riguarda
  - la comprensione e la lettura delle sorgenti informative
  - la copiatura nell'area di preparazione dei dati delle porzioni di sorgenti informative che sono necessarie al popolamento del data warehouse

## Trasformazione

I dati estratti dalle sorgenti informative, prima di essere caricati nel data warehouse, sono sottoposti a diverse **trasformazioni**

- pulizia
  - per risolvere errori, conflitti, incompletezze
  - per riportare i dati in un formato standard
- eliminazione di campi non significativi
- combinazione
  - per identificare e correlare i dati associati alla rappresentazione di uno stesso oggetto in più sorgenti informative
- creazione di chiavi
  - le chiavi usate nel data warehouse sono normalmente diverse da quelle usate nelle sorgenti informative
- creazione di aggregati

## Caricamento e controllo di qualità

Dopo estrazione e trasformazione, i dati sono organizzati per essere caricati direttamente nel data warehouse

- il caricamento consiste nella concatenazione (e/o aggiornamento) di un insieme di record per ciascuna tabella (fatti o dimensione) del data warehouse
  - durante il caricamento il data warehouse non è solitamente disponibile per l'accesso e l'interrogazione
- il caricamento dei dati nel data warehouse viene seguito da una verifica della correttezza delle operazioni di preparazione e caricamento, mediante un'analisi di qualità dei dati
  - se il controllo di qualità ha successo, il nuovo data warehouse è pronto per l'accesso e l'interrogazione

## Aggiornamento del data warehouse

I dati del data warehouse devono essere aggiornati, anche frequentemente

- aggiornamenti ordinari e periodici
  - caricamento incrementale di nuovi dati nel data warehouse
- aggiornamenti straordinari
  - correzione di dati (record e/o schemi)
  - sono aggiornamenti orientati al miglioramento della qualità complessiva dei dati

## Interrogazione del data warehouse

L'interrogazione (o analisi) è l'attività prevalente nel data warehouse

- il data warehouse è stato creato per essere interrogato
- il data warehouse deve essere ottimizzato per l'esecuzione di interrogazioni complesse
  - ad esempio, mediante la gestione (trasparente) di dati aggregati
- l'interrogazione avviene mediante diversi strumenti

## Attività di amministrazione

### Auditing

- sull'origine dei dati (ad esempio, per certificarne la qualità)
- sull'uso del data warehouse (per l'ottimizzazione del data warehouse)

### Gestione della sicurezza

### Backup e recovery

### Gestione del feedback

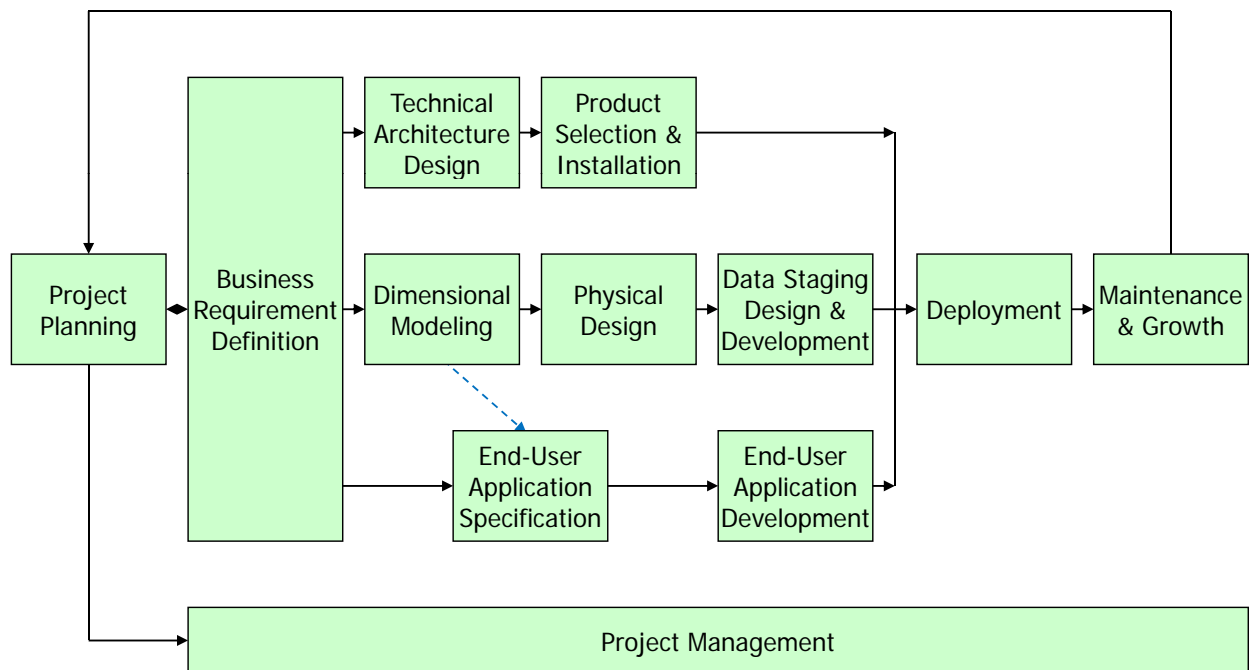
- il transito principale dei dati va dalle sorgenti informative al data warehouse e dal data warehouse agli strumenti di analisi
  - dati "puliti" e risultati di analisi significativi possono transitare nella direzione opposta

## Ciclo di vita dimensionale

Il **ciclo di vita dimensionale** (**Business Dimensional 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
- mediante un insieme di attività e di relazioni tra attività

## Ciclo di vita dimensionale



## Fasi nel ciclo di vita dimensionale

- pianificazione del progetto
- gestione del progetto
- raccolta e analisi dei requisiti
- progettazione del data warehouse
  - progettazione dei dati
    - progettazione dimensionale, progettazione fisica, progetto della preparazione dei dati
  - progettazione tecnologica
    - progettazione dell'architettura tecnica, selezione e installazione dei prodotti
  - progettazione delle applicazioni
    - specifica delle applicazioni, sviluppo delle applicazioni
- installazione e avviamento
- manutenzione e crescita