Data Warehousing, Data Mining & Business Intelligence

DataWarehousing

Paolo G. Franciosa

Dipartimento di Statistica, Probabilità e Statistiche Applicate

Università "La Sapienza"

paolo.franciosa@uniroma1.it

Questo materiale deriva dalla traduzione e adattatamento delle presentazioni pubblicate dal prof. Jiawei Han

Data Warehousing

- Cosa è un Data Warehouse?
- Il modello di dati multidimensionale
- Architettura di un Data Warehouse
- Realizzazione di un Data Warehouse
- Evoluzione dei Data Cube

Cosa è un Data Warehouse?

- Un DB separato dal DB operazionale
- Contiene informazioni, anche di carattere storico, utili per scopi di analisi
- Le informazioni sono organizzate per argomenti, eliminando dati non utili
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process." [W. H. Inmon]
- Data warehousing: progettazione, realizzazione e popolamento di un DW

Data Warehouse: Subject-Oriented

- Organizzato per scopi particolari (ev. Data Marts)
- Non è necessario mantenere informazioni non rilevanti per il particolare obiettivo

Data Warehouse: Integrato

- Informazioni estratte da diverse fonti (DB, files, altro)

- I dati sono soggetti a operazioni di ripulitura, filtraggio e allineamento (divergenze sui nomi, omonimie, unità di misura, codifiche, …)

Data Warehouse: Time Variant

- L'orizzonte temporale di un DW è decisamente più lungo di quello dei DB operazionali
  - DB Operazionali: stato corrente
  - DW: prospettiva storica

- I dati utilizzati per attività di OLAP/DM non sono necessariamente aggiornati in tempo reale

Data Warehouse: Non-Volatile

- Il DW è fisicamente separato dal DB operazionale
- Il DW non è soggetto a modifiche on-line.
  - Non richiede meccanismi di gestione delle transazioni, recupero della consistenza, controllo della concorrenza
- Prevede solo due tipi di operazioni: caricamento dei dati e accesso ai dati

Separare DB operazionali da DW

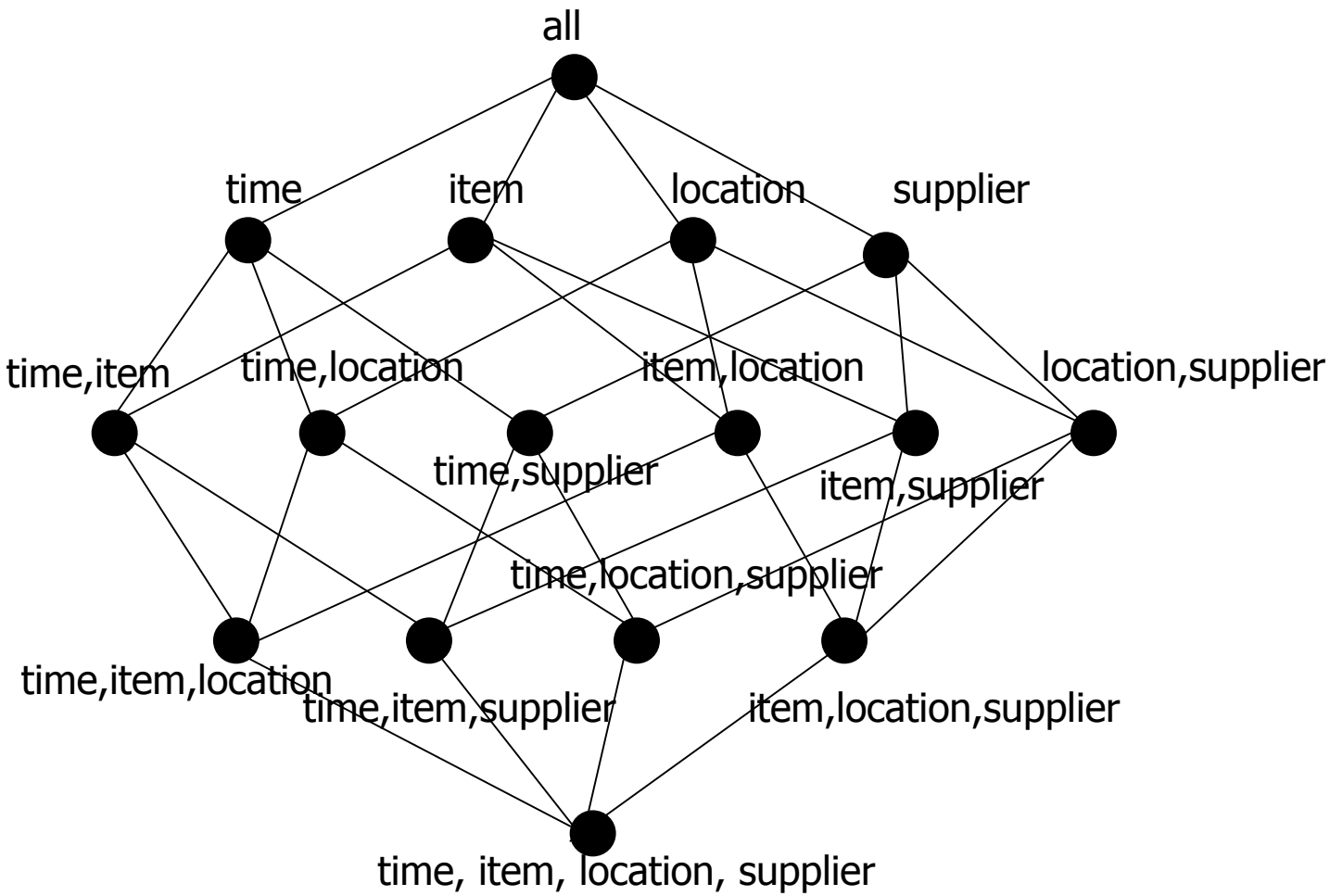
- Massimizzare le prestazioni (transazioni-robustezza vs. scansioni-persistenza)
- Differente trattamento dei dati:
  - dati mancanti: manteniamo stati precedenti
  - dati consolidati: il DW mantiene esplicitamente valori aggregati per migliorare le prestazioni
  - qualità dei dati: I DB operazionali possono essere diversi, il DW può prevedere trasformazioni dei dati dovuti alla riconciliazione, cleaning e allineamento

Il modello multidimensionale

- I dati del DW sono organizzati e visualizzati in base alle dimensioni e alla granularità delle aggregazioni (per ogni dimensione)
  - dimensioni (livelli di granularità): prodotto(nome, marca, categoria), o tempo(giorno, settimana, mese, stagione, anno)
  - *fact tables*: contiene misure associate a riferimenti a dimensioni associati a una k-upla delle dimensioni ad una certa granularità, p.es.
    - (fatturato, num_ordini, detersivi, gennaio-2000)
    - (fatturato, num_ordini, brillatutto, 2001)

- Il cubo di base n-dimensionale è chiamato base cuboid. Il cubo più in alto nella gerarchia (al massimo livello di aggregazione) è chiamato apex cuboid. Il *lattice* dei cuboids è il data cube.

Struttura del Data Cube (*lattice*)



0-D(apex) cuboid

1-D cuboids

2-D cuboids

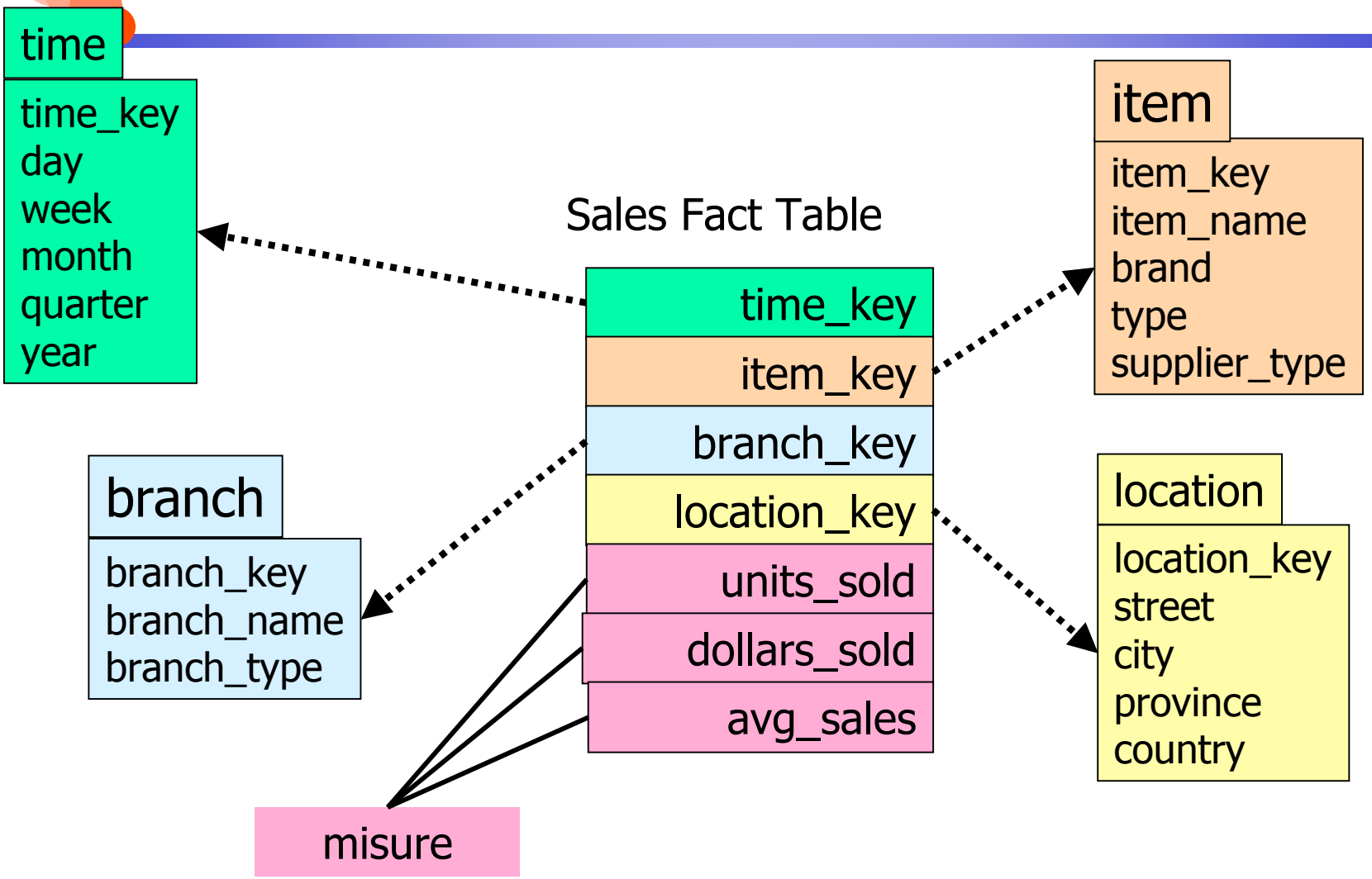
3-D cuboids

4-D(base) cuboid

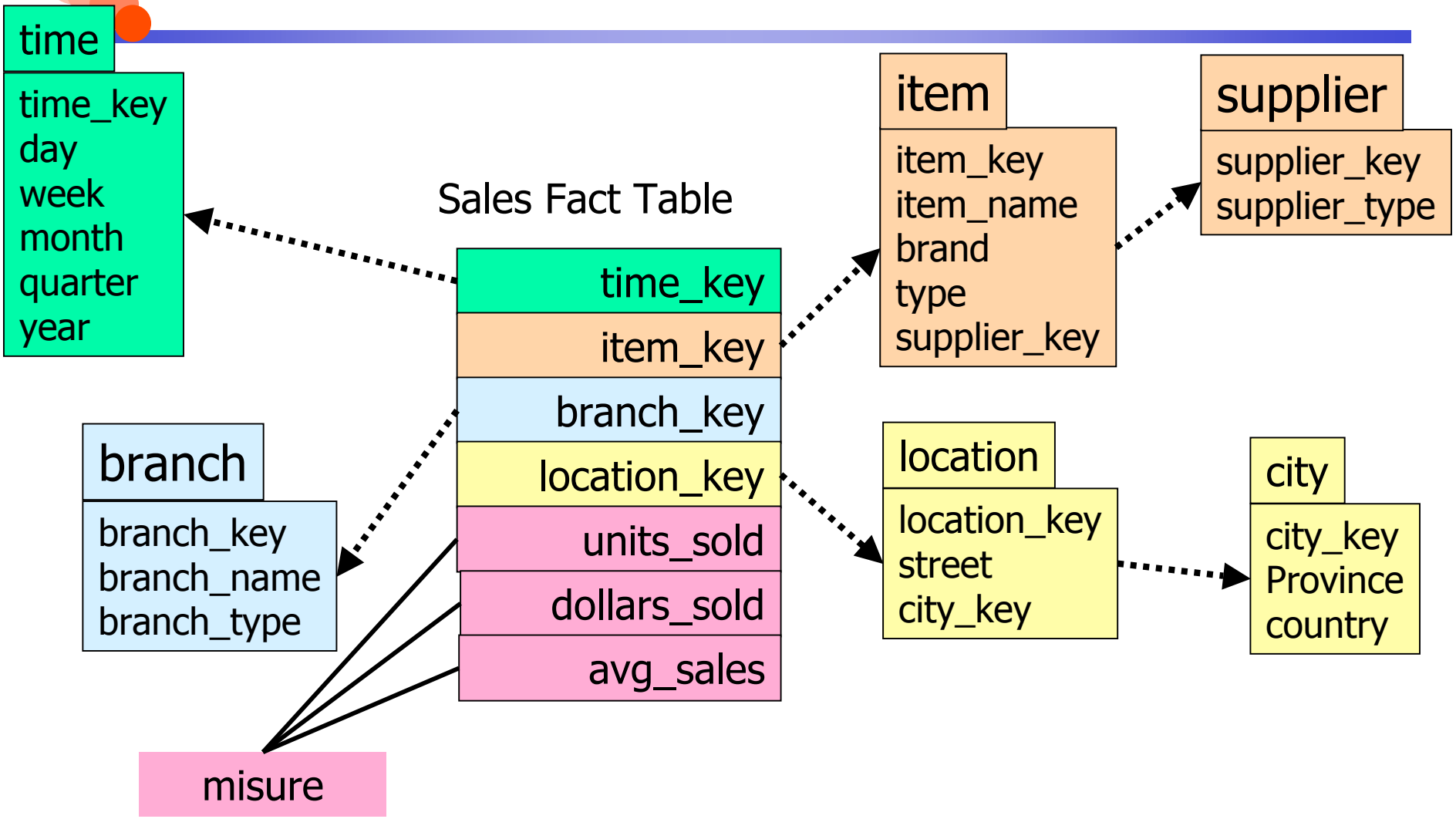
Modello concettuale di un Data Warehouse

- Dimensioni e misure
  - Star schema: una fact table al centro, connessa a varie dimension tables
  - Snowflake schema: un raffinamento dello star schema, in cui alcune gerarchie di dimensioni sono normalizzate in un insieme di altre dimension tables
  - Fact constellations: più fact tables che possono condividere alcune dimension tables, chiamato anche *galaxy schema*

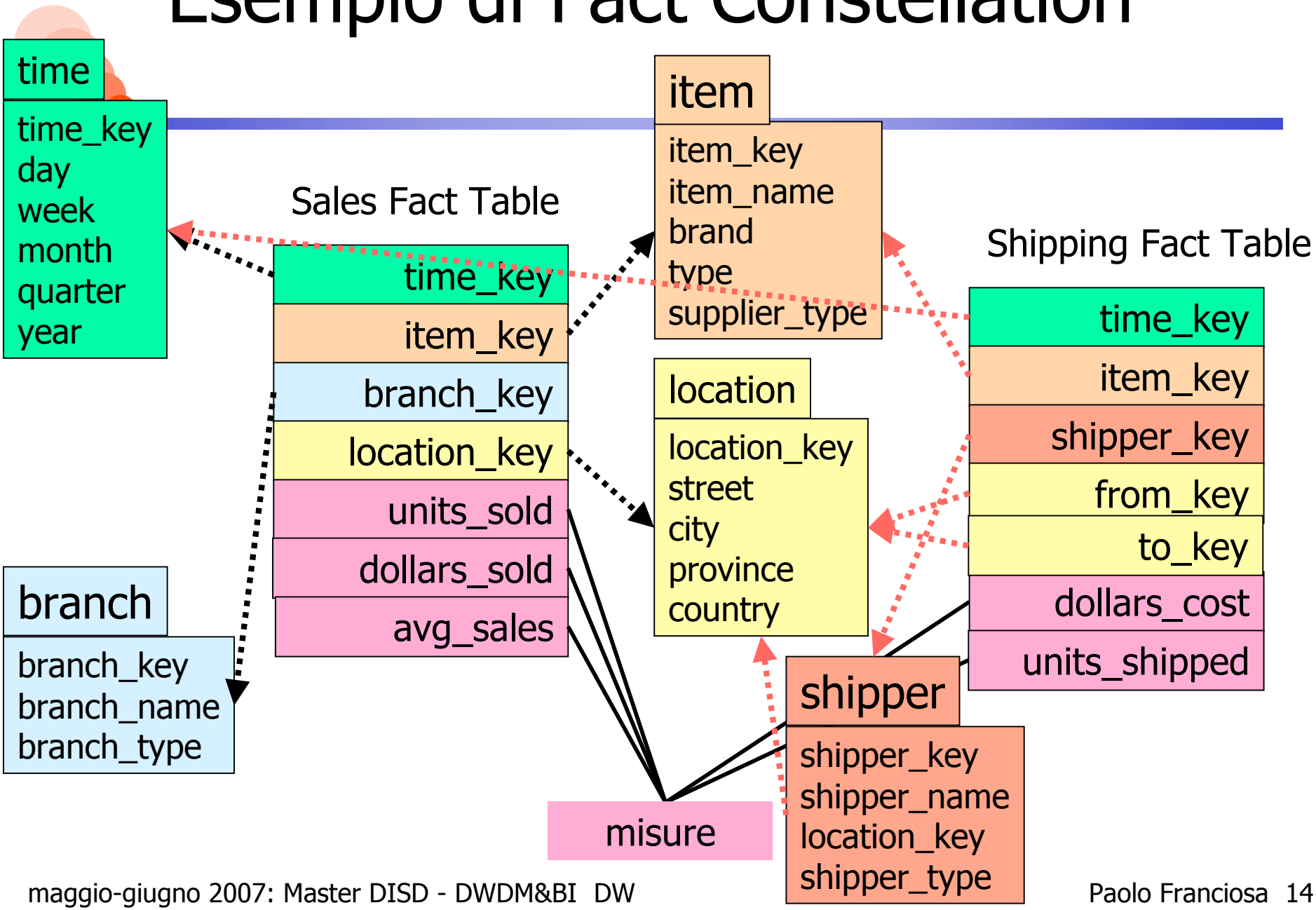
Esempio di Star Schema



Esempio di Snowflake Schema



Esempio di Fact Constellation



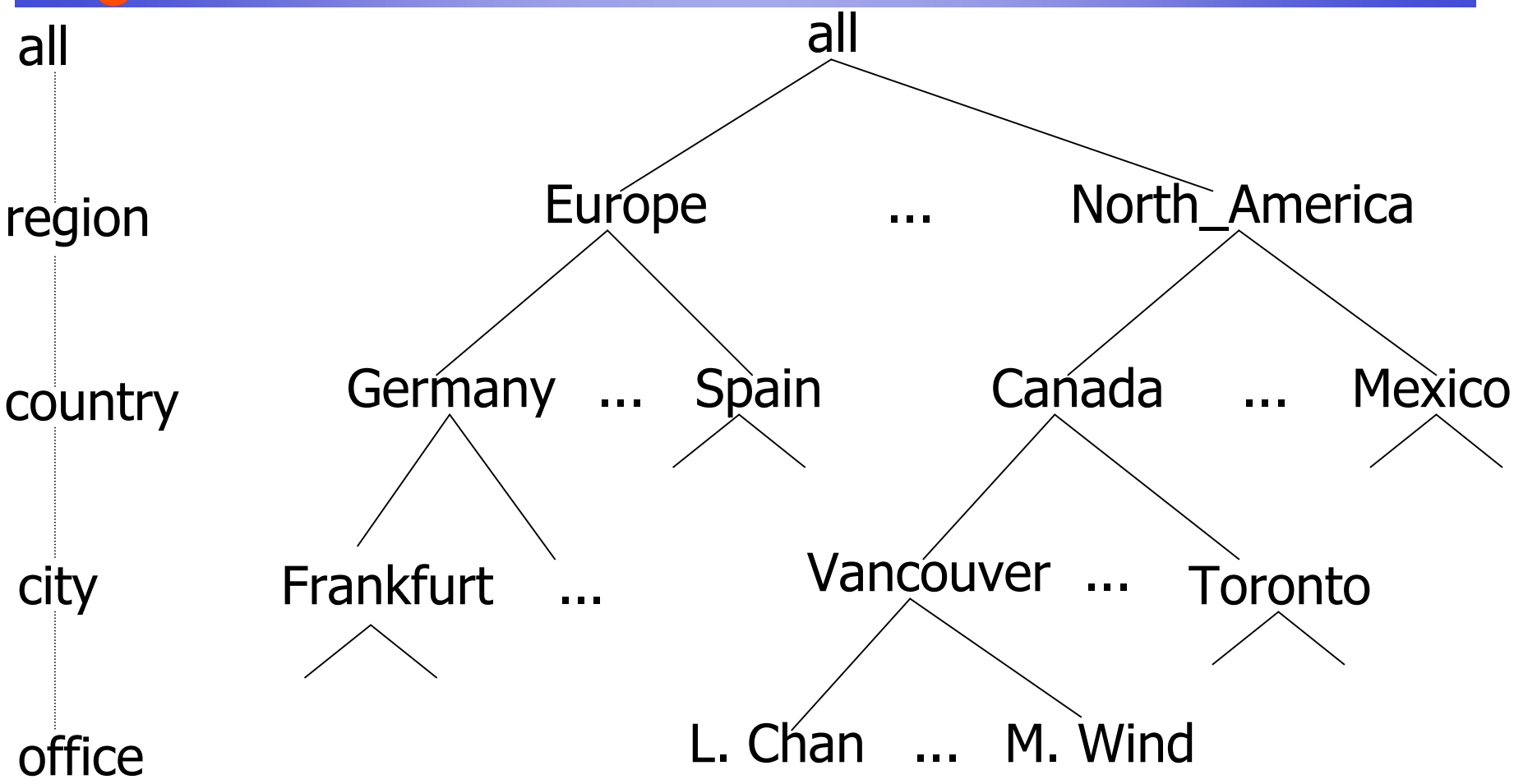
Misure

- Gli attributi delle fact table sono dimensioni e misure
- Il valore di una misura dipende dal livello di aggregazione (nella gerarchia delle dimensioni)
- Le prestazioni di uno strumento OLAP possono migliorare se le misure sono rappresentate in modo opportuno

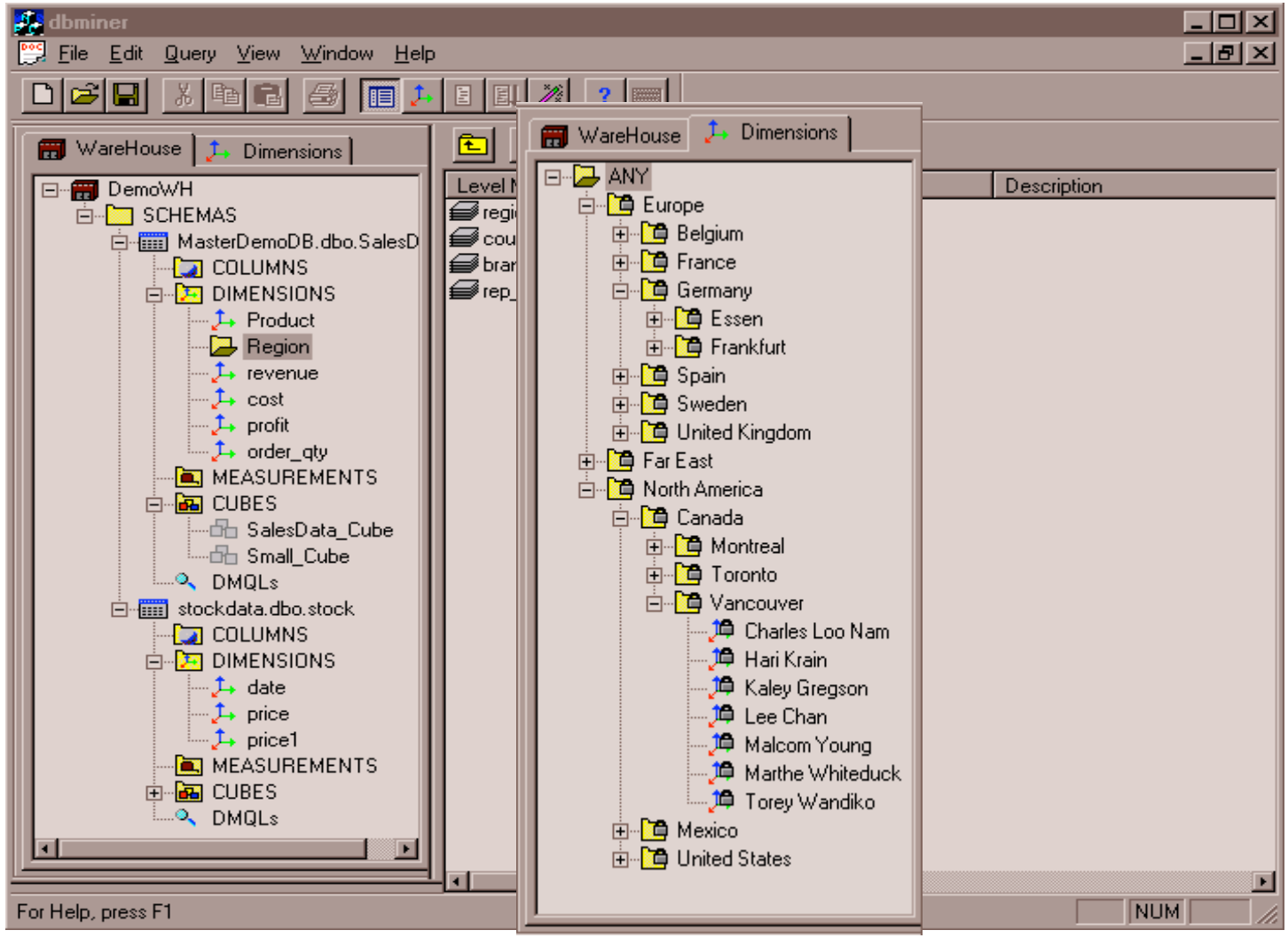
Funzioni di aggregazione

- distributive: se il risultato non cambia applicando livelli intermedi di aggregazione
  - p.es. min(a, b, c, d, e) = min(min(a, b, c), min(d, e))
  - count(), sum(), min(), max().
- algebriche: se può essere calcolata attraverso una funzione algebrica con un numero costante di argomenti, ciascuno dei quali è ottenuto applicando una funzione distributiva
  - p.es. avg(a, b, c, d, e) =
    (sum(a, b, c) + sum(d, e)) / (count(a, b, c) + count(d, e))
- globali: non si può ricavare a partire da un numero costante di misure sui sottoinsiemi
  - E.g., median(), mode(), rank().

Un esempio di dimensione: località



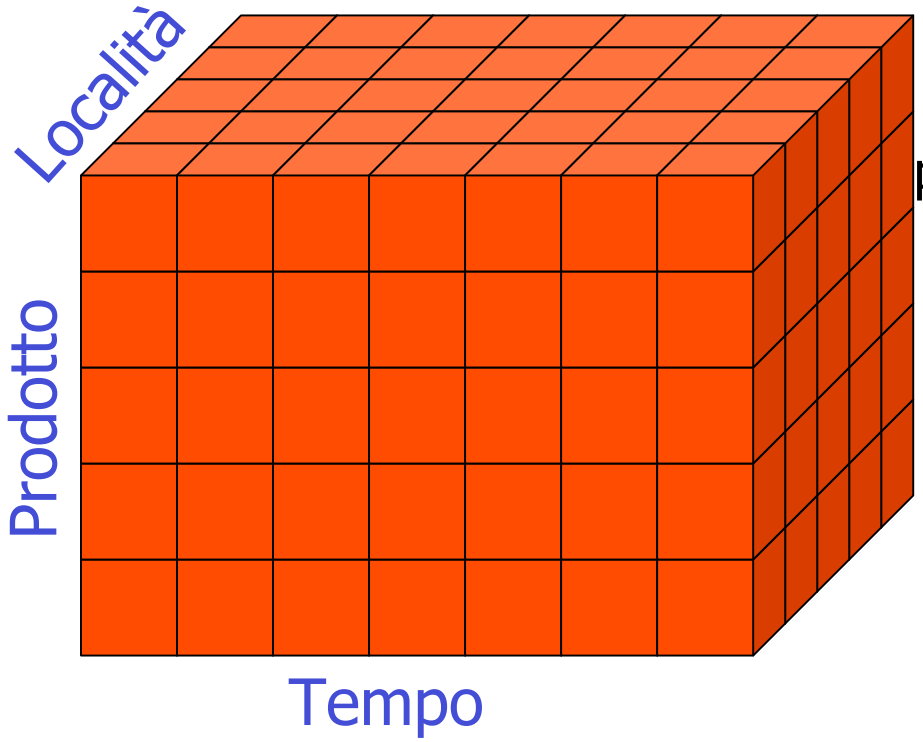
Descrizione dello schema di un DW



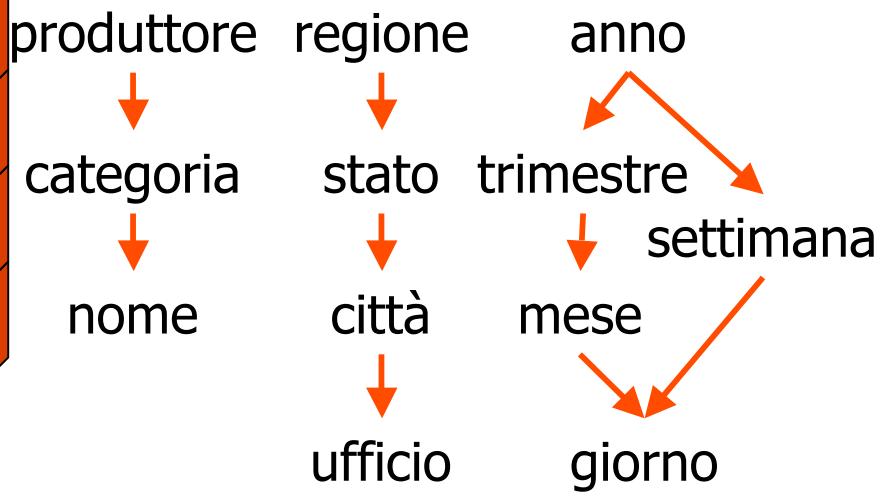
Dati Multidimensionali

- Vendite in funzione del prodotto, periodo e località

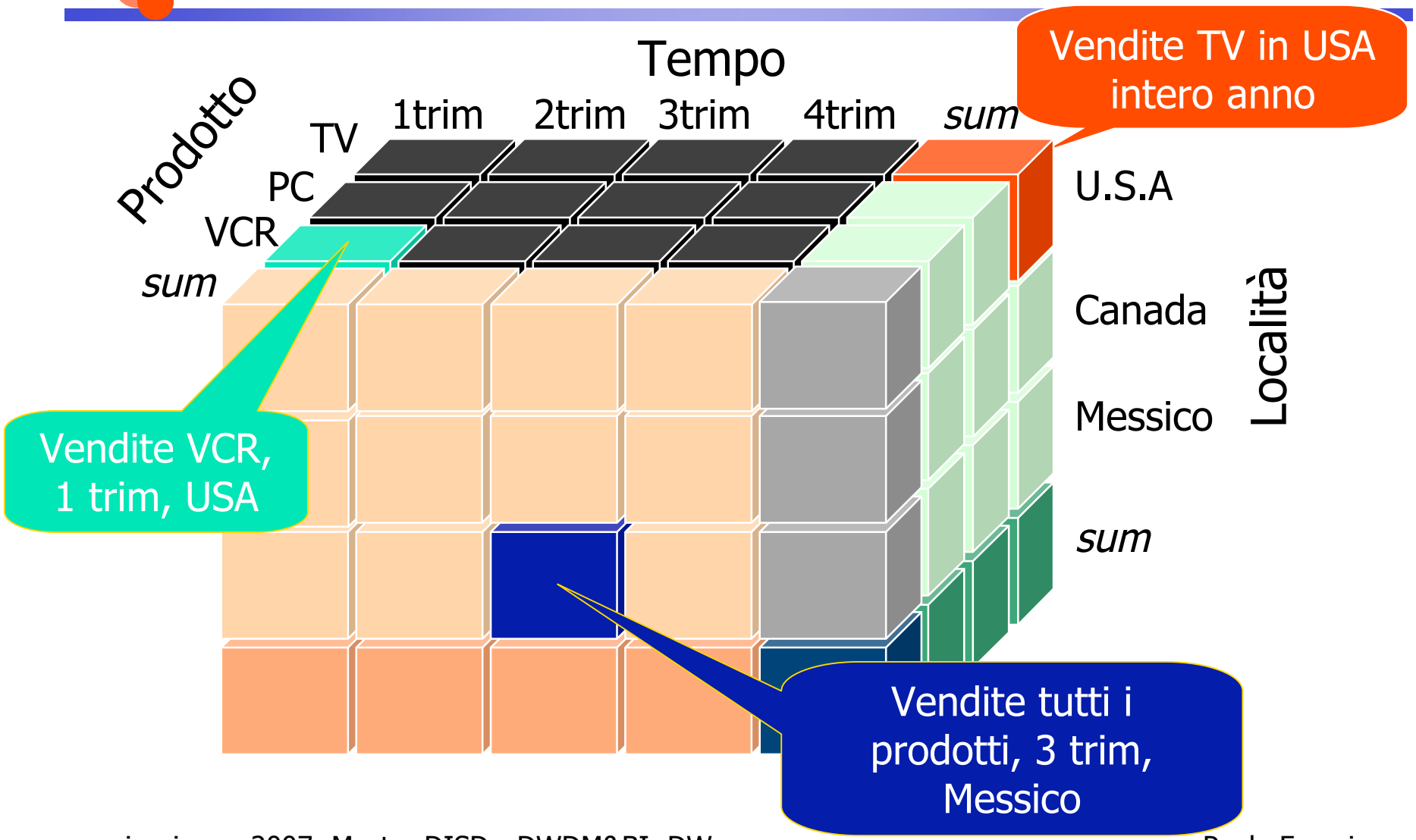
Dimensioni: Prodotto, Località, Tempo



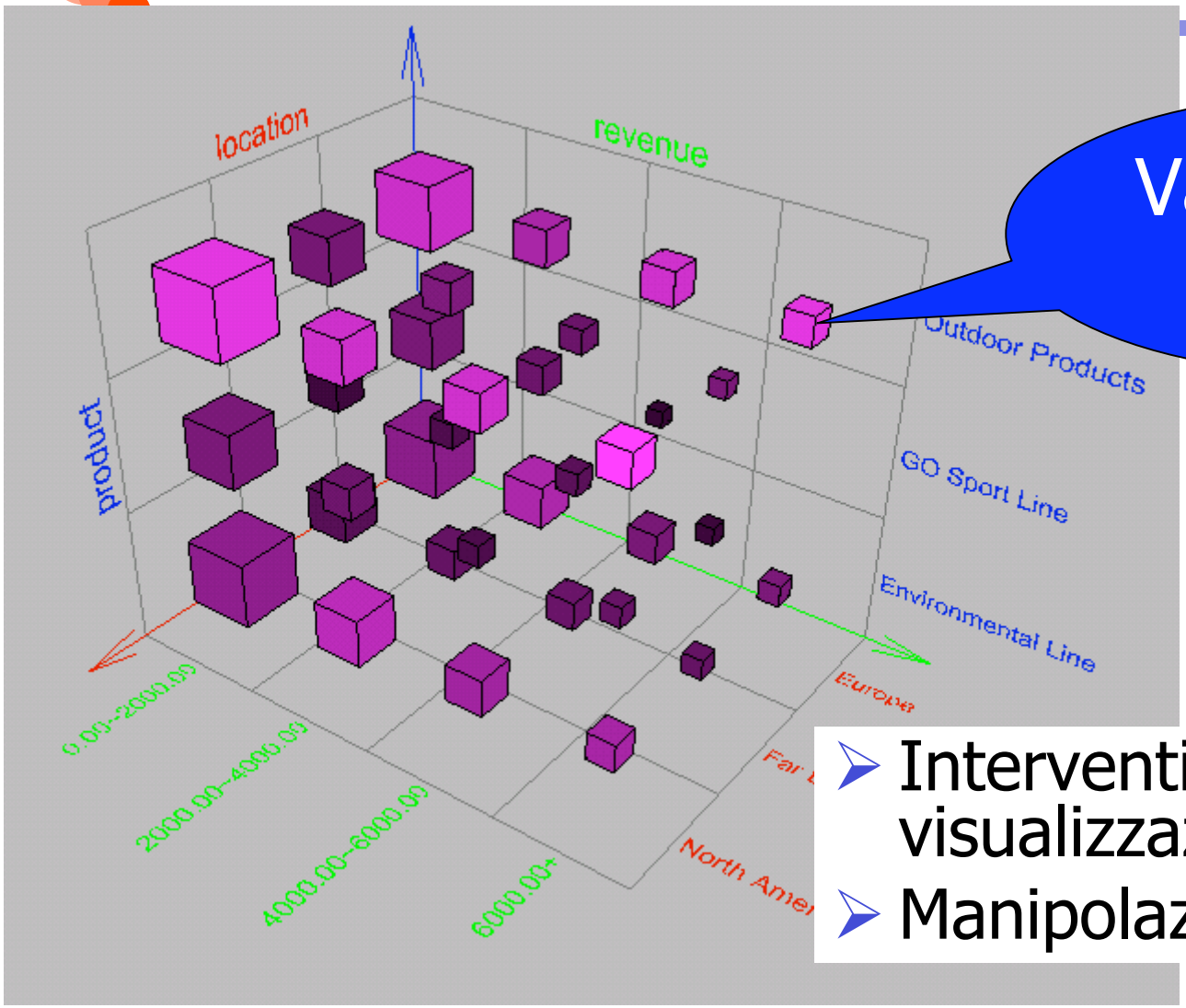
Percorsi di aggregazione:



Visualizzazione di un Data Cube



Navigazione in un Data Cube



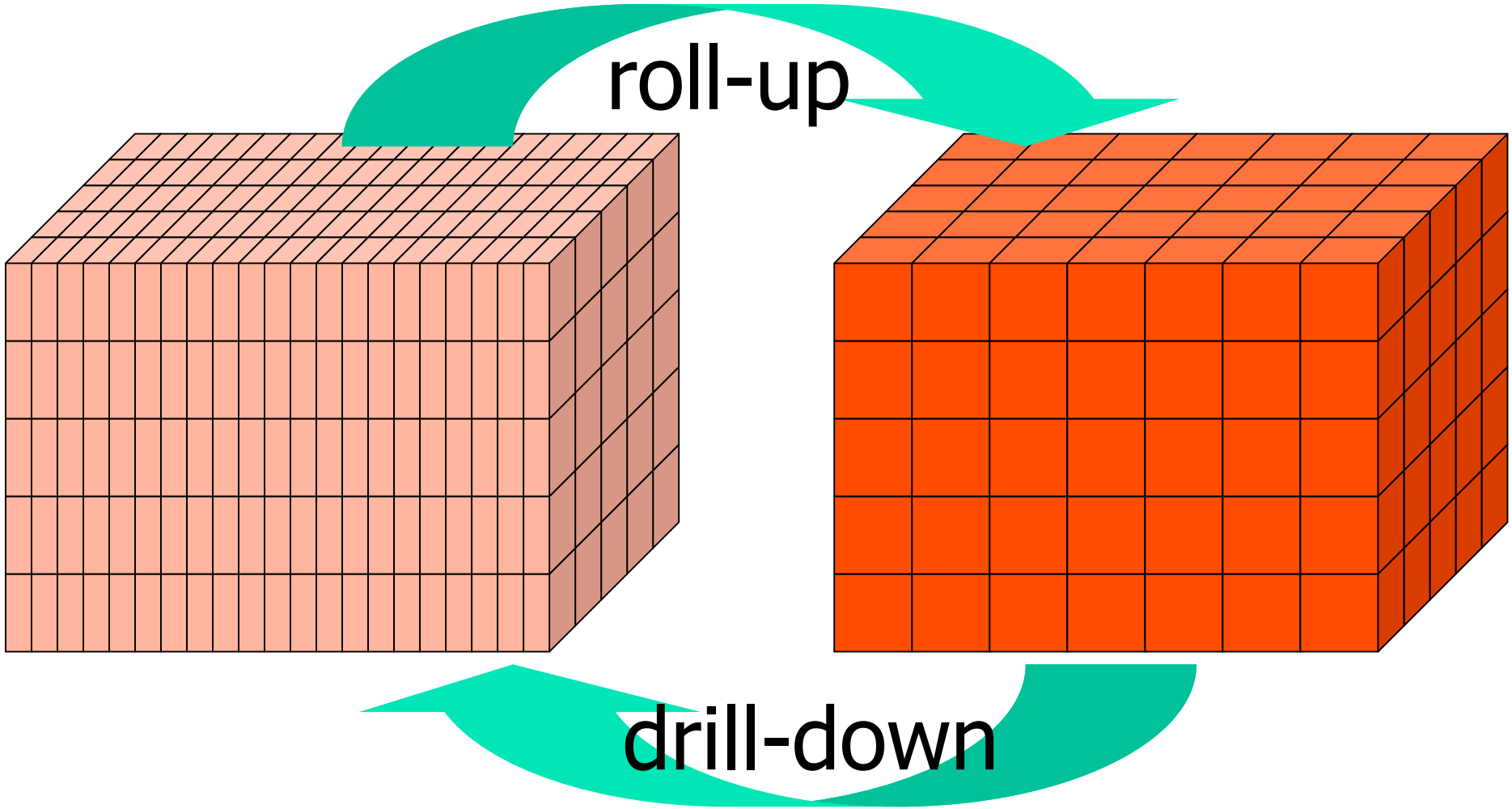
Valore della misura

- Interventi sulla visualizzazione
- Manipolazione interattiva

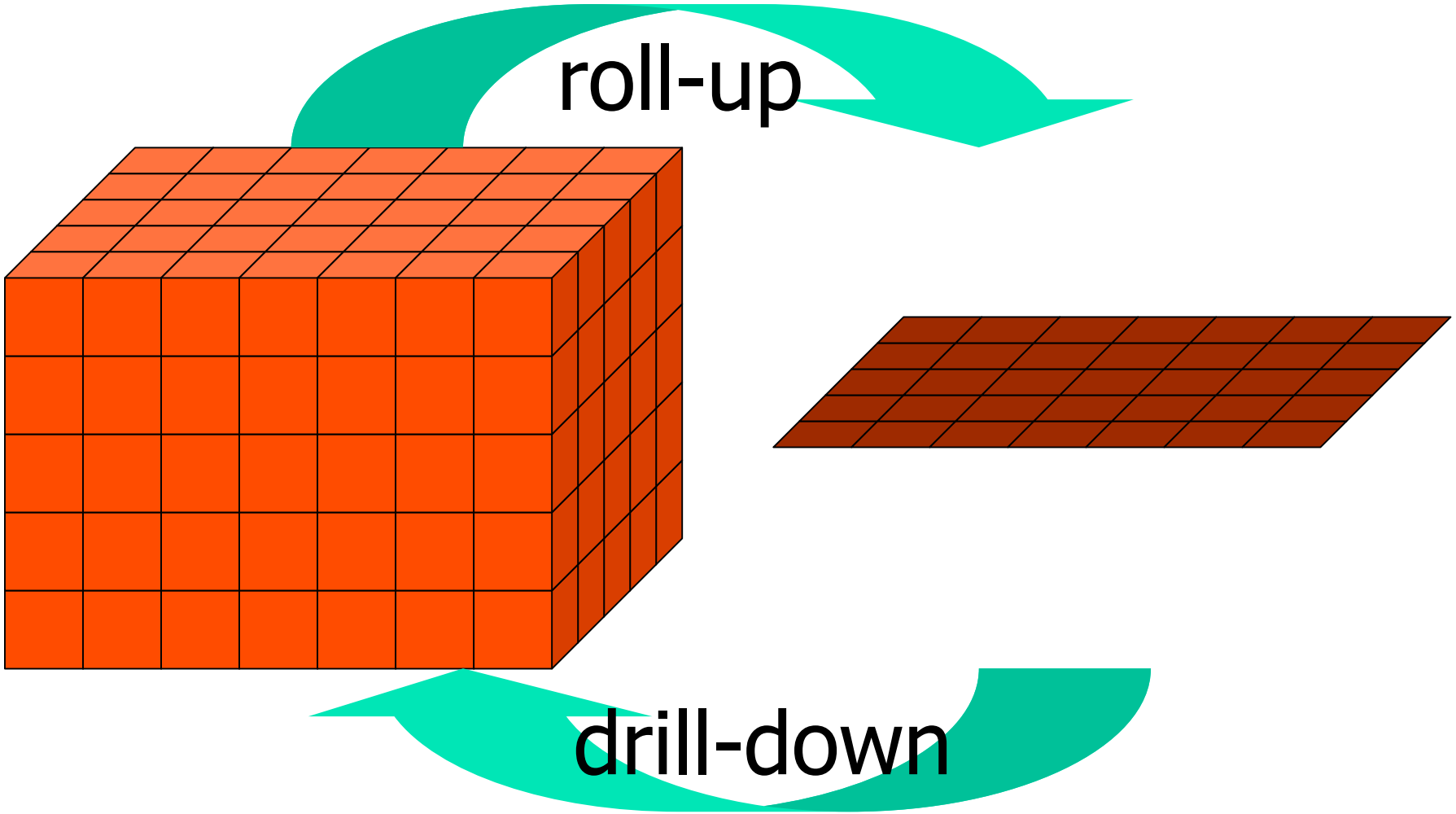
Operazioni OLAP tipiche

- Roll up (drill-up): aggrega dati
  - *Risalire nella gerarchia di una dimensione (minor livello di dettaglio)*
  - *Proiettare su alcune dimensioni (elimina le altre)*
- Drill down (roll down): inversa del roll-up
  - *Scendere nella gerarchia di una dimensione (maggior livello di dettaglio)*
  - *Aggiungere nuove dimensioni*
  - *drill through: verso il "base cube", interroga i microdati (SQL)*
- Slice and dice:
  - *project and select*
- Pivot (rotate): primitiva di visualizzazione
  - *Rotazioni, trasformazione da 3D a una serie di piani 2D.*

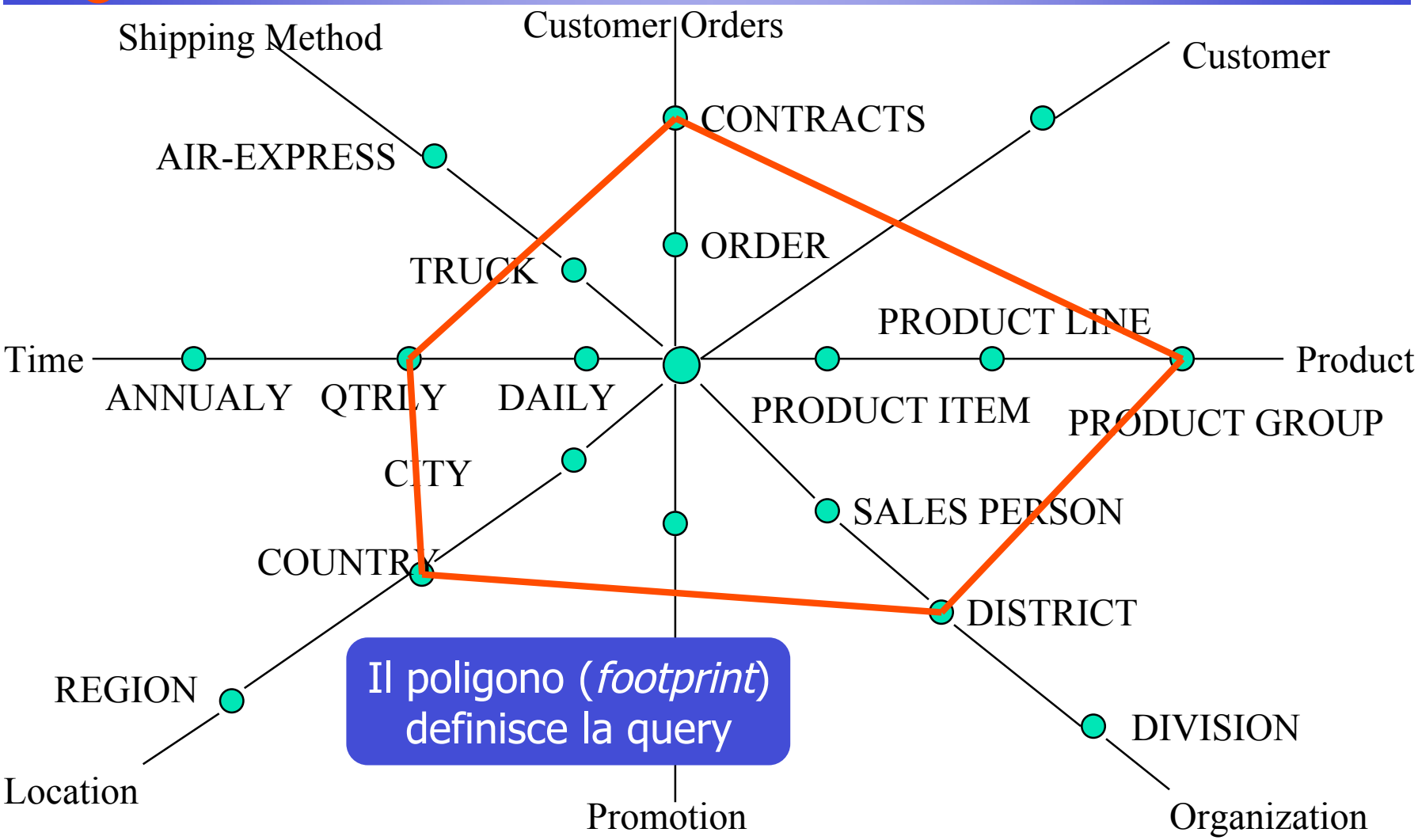
Modificare livello di dettaglio



Proiettare – Aggiungere dimensioni



Una modalità di specifica delle query



Il poligono (*footprint*) definisce la query

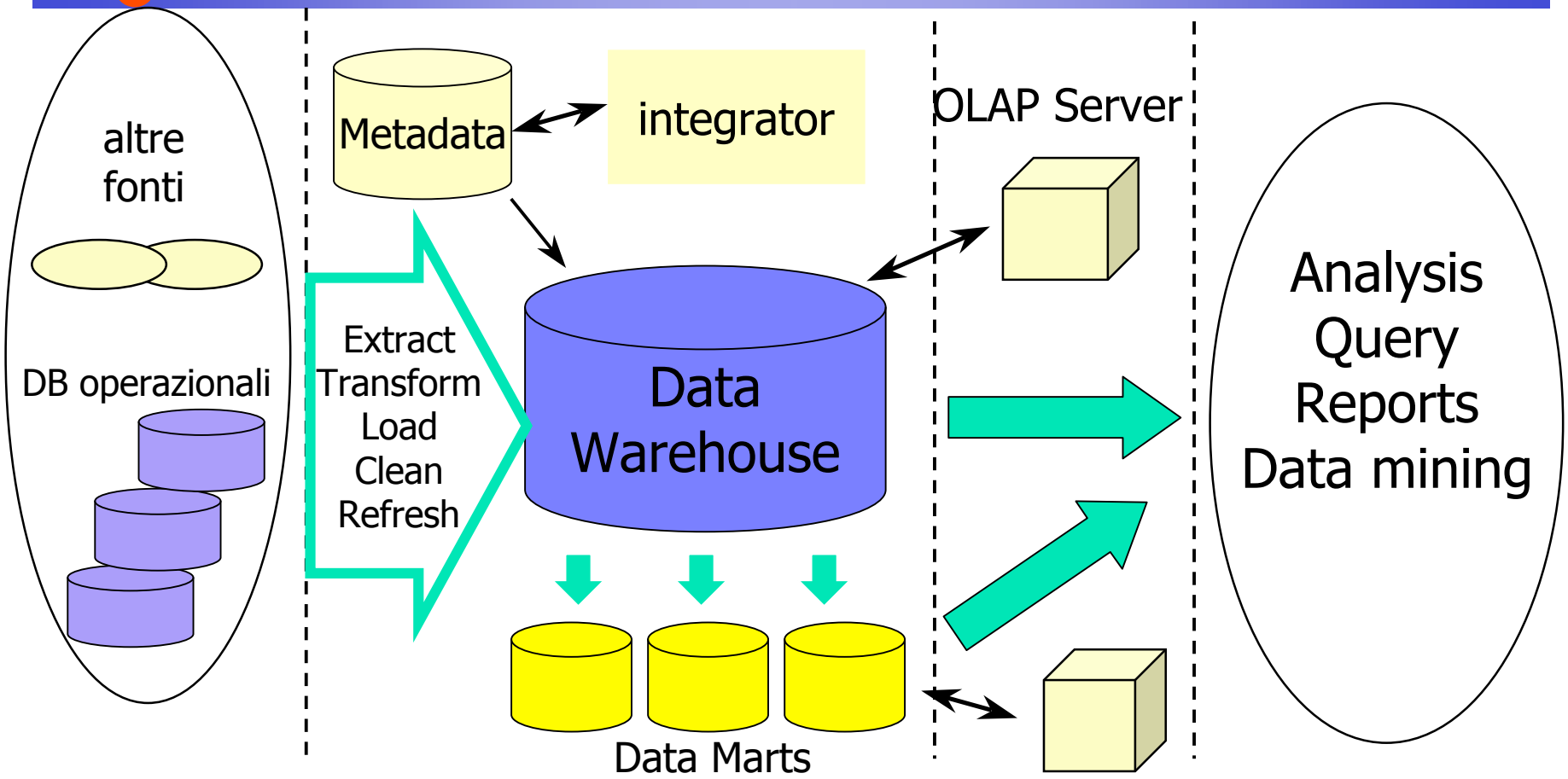
Progettazione di un DW

- Top-down: progetto globale (adatto a contesti consolidati)
- Bottom-up: prototipazione e sperimentazione immediata
- Misto

- Nel linguaggio dell'ingegneria del software:
  - Sviluppo a cascata: successione ben definita di passi, ciascuno completato con un collaudo del prodotto intermedio
  - Sviluppo a spirale: una successione di sistemi completi, via via arricchiti di funzionalità

- Scelte nella progettazione di un Data Warehouse
  - processo di business da modellare: p.es. ordini, fatture, etc.
  - *granularità* (livello minimo di dettaglio) dei dati di business
  - dimensioni da applicare a ciascun record delle fact table
  - misure associate a ciascun record delle fact table

Una architettura a più livelli



Data Sources Data Storage OLAP Engine Front-End Tools

DW a 1–2–(0) Livelli

- Enterprise warehouse
  - è il repository dell'intera azienza, comune a tutte le attività di business
- Data Marts
  - sono porzioni estratte dal Data Warehouse, destinate a particolari gruppi di utenti. Ciascun data mart contiene solo le informazioni utili ad una certa categoria di analisi/indagini
  - possono essere materializzati
- Virtual warehouse
  - un insieme di viste sui DB operazionali
  - può essere materializzato
  - attuabile solo in presenza di un DB operazionale che non richieda attività profonde di cleaning, trasformazione, ...

Metadata Repository

I metadati sono i dati che descrivono il Data Warehouse

- struttura del data warehouse
  - Schema, viste, dimensioni e loro gerarchie, regole di derivazione, definizione dei data mart
- metadati operazionali
  - provenienza dei dati (data lineage, history of migrated data and transformation path), limiti di validità dei dati, informazioni sull'uso del DW (statistiche, logs, error reports, ...)
- algoritmi usati per l'aggregazione
- Procedure di importazione dal DB operazionale al DW
- amministrazione e utenti del DW

Data Warehouse Back-End Tools

- Estrazione
  - preleva i dati da sorgenti eterogenee
- Data cleaning
  - individua errori nei dati importati e, se possibile, li rettifica
- Data transformation
  - converte i dati dal formato di provenienza a quello previsto per il DW
- Load
  - ordina, calcola aggregazioni, verifica vincoli di integrità, costruisce indici, ...
- Refresh
  - propaga periodicamente gli aggiornamenti dalle sorgenti informative al DW

Uso del Data Warehouse

- Tre tipi di utilizzo
  - Information processing
    - interrogazioni, statistiche di base, semplici report/grafici
  - OLAP
    - analisi multidimensionale
    - operazioni di base: slice&dice, drilling, pivoting
  - Data Mining
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.