Università degli Studi di Palermo

dipartimento di ingegneria unipa

Data Warehousing e Data Lake

CORSO DI BIG DATA

a.a. 2022/2023

Prof. Roberto Pirrone

LABORATORIO DI INTERAZIONE UOMO-MACCHINA
CHILAB

Sommario

- Introduzione
  - Basi di dati integrate, sì, ma …
  - OLTP e OLAP
- Data warehouse e data warehousing
- Dati multidimensionali
- Data Lake

Base di dati

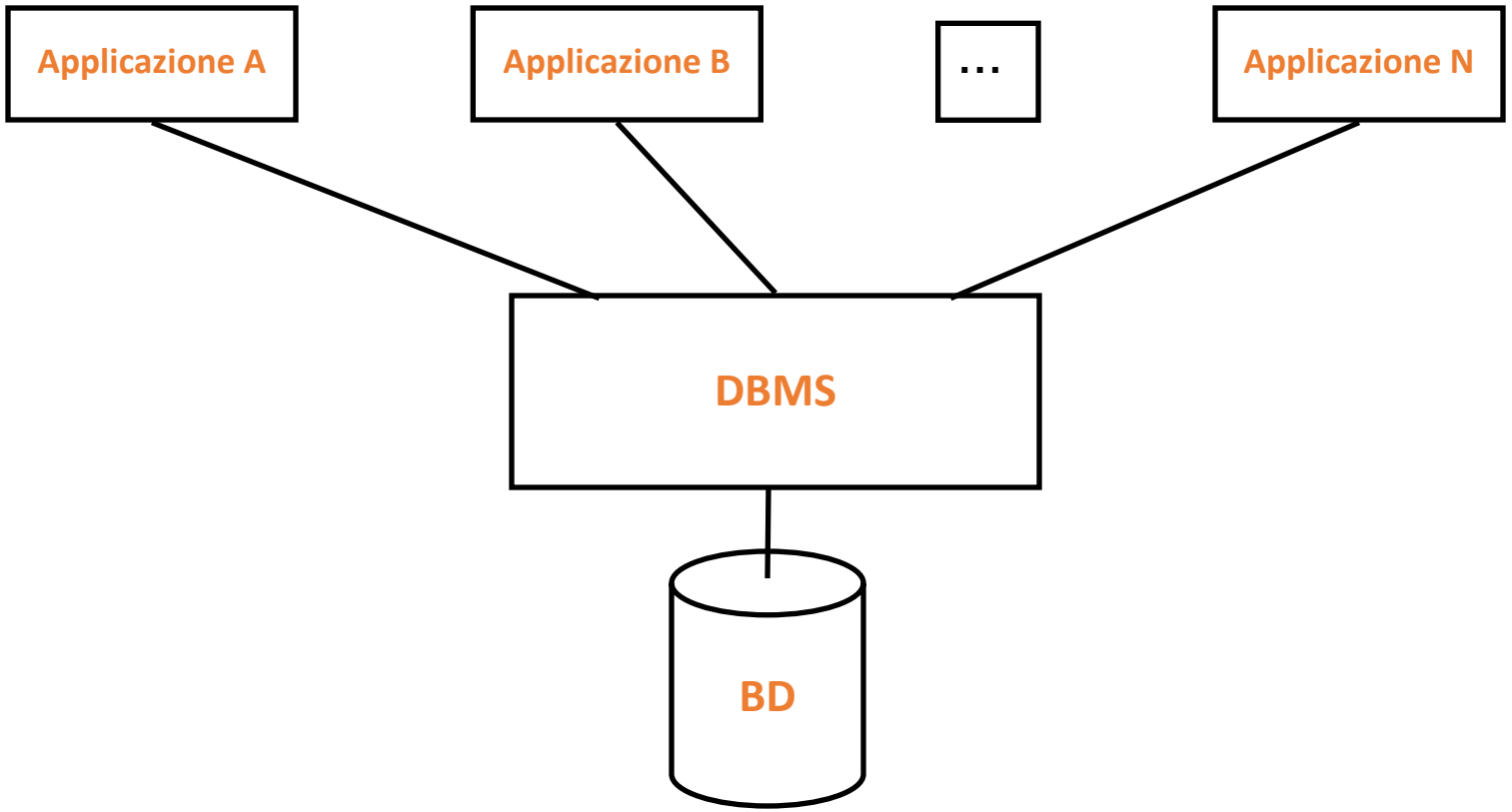
- “Collezione di dati **persistente** e **condivisa**, gestita in modo **efficace**, **efficiente** e **affidabile** (da un **DBMS**)”

- il concetto di base di dati nasce per rispondere alle esigenze di “gestione di una risorsa pregiata”, *condivisa* da più applicazioni

Base di dati «ideale»

- “ogni organizzazione ha *una* base di dati, che organizza tutti i dati di interesse in forma integrata e non ridondante”
- “ciascuna applicazione ha accesso a tutti i dati di proprio interesse, in tempo reale e senza duplicazione, riorganizzati secondo le proprie necessità”
- …

Base di dati «ideale»



L’obiettivo ideale è sensato e praticabile?

- La realtà è in continua evoluzione, non esiste uno “stato stazionario” (se non nell’iperuranio):
  - cambiano le esigenze
  - cambiano le strutture
  - le realizzazioni richiedono tempo
- Il coordinamento forte fra i vari settori può risultare controproducente
- Ogni organizzazione ha di solito diverse basi di dati *distribuite, eterogenee, autonome*
- Ad esempio, la nostra università:
  - … quali sistemi e basi di dati ha (per quanto ne sappiamo)?
  - … proviamo a pensarci

Risorse e Processi

- *Risorsa*
  - tutto ciò con cui l'organizzazione opera, sia materiale che immateriale, per perseguire i suoi obiettivi
    - le informazioni, i dati sono risorse
- *Processo*
  - l'insieme di attività (sequenze di decisioni e azioni) che l'organizzazione nel suo complesso svolge per raggiungere un obiettivo, gestendo il ciclo di vita di una risorsa o di un gruppo omogeneo di risorse

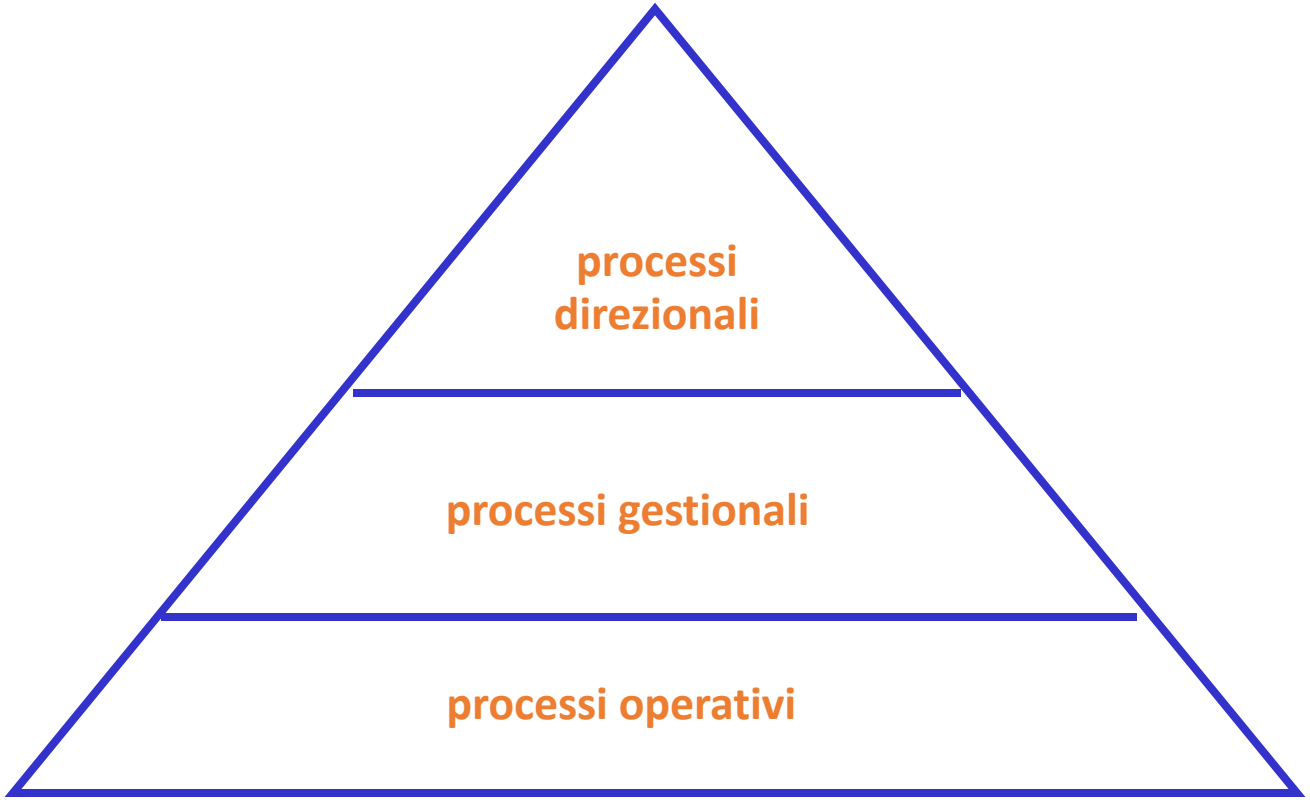
Processi presso una banca

- gestione di un movimento su un conto corrente bancario, presso sportello tradizionale o automatico

- concessione di un fido
- revisione delle condizioni su un conto corrente

- verifica dell’andamento dei servizi di carta di credito
- lancio di una campagna promozionale
- stipula di accordi commerciali
- fusione con un'altra banca

Processi


processi direzionali
processi gestionali
processi operativi

Processi presso una banca

- *Processi operativi*
  - gestione di un movimento su un conto corrente bancario, presso sportello tradizionale o automatico
- *Processi gestionali*
  - concessione di un fido
  - revisione delle condizioni su un conto corrente
- *Processi direzionali*
  - verifica dell'andamento dei servizi di carta di credito
  - lancio di una campagna promozionale
  - stipula di accordi commerciali

Processi presso un’azienda telefonica

- *Processi operativi*
  - stipula di contratti ordinari
  - instradamento delle telefonate
  - memorizzazione di dati contabili sulle telefonate (chiamante, chiamato, giorno, ora, durata, instradamento,..)
- *Processi gestionali*
  - stipula di contratti speciali
  - installazione di infrastrutture
- *Processi direzionali*
  - scelta dei parametri che fissano il costo delle telefonate
  - definizione di contratti diversificati
  - pianificazione del potenziamento delle infrastrutture

Processi presso l'università

- *Processi operativi*
- *Processi gestionali*
- *Processi direzionali*

Caratteristiche dei processi dei vari tipi

- Processi operativi
  - su dati dipartimentali e dettagliati
  - operazioni strutturate, basate su regole perfettamente definite
- Processi gestionali
  - su dati settoriali e parzialmente aggregati
  - operazioni semi-strutturate, basate su regole note, ma con un intervento umano con assunzione di responsabilità
- Processi direzionali
  - su dati integrati e fortemente aggregati
  - operazioni non strutturate, senza criteri precisi: capacità personale è essenziale

Sistemi informatici: una classificazione

- per i processi operativi
  - *Transaction processing systems*
- per i processi gestionali
  - *Management information systems* (di solito settoriali)
- per i processi direzionali

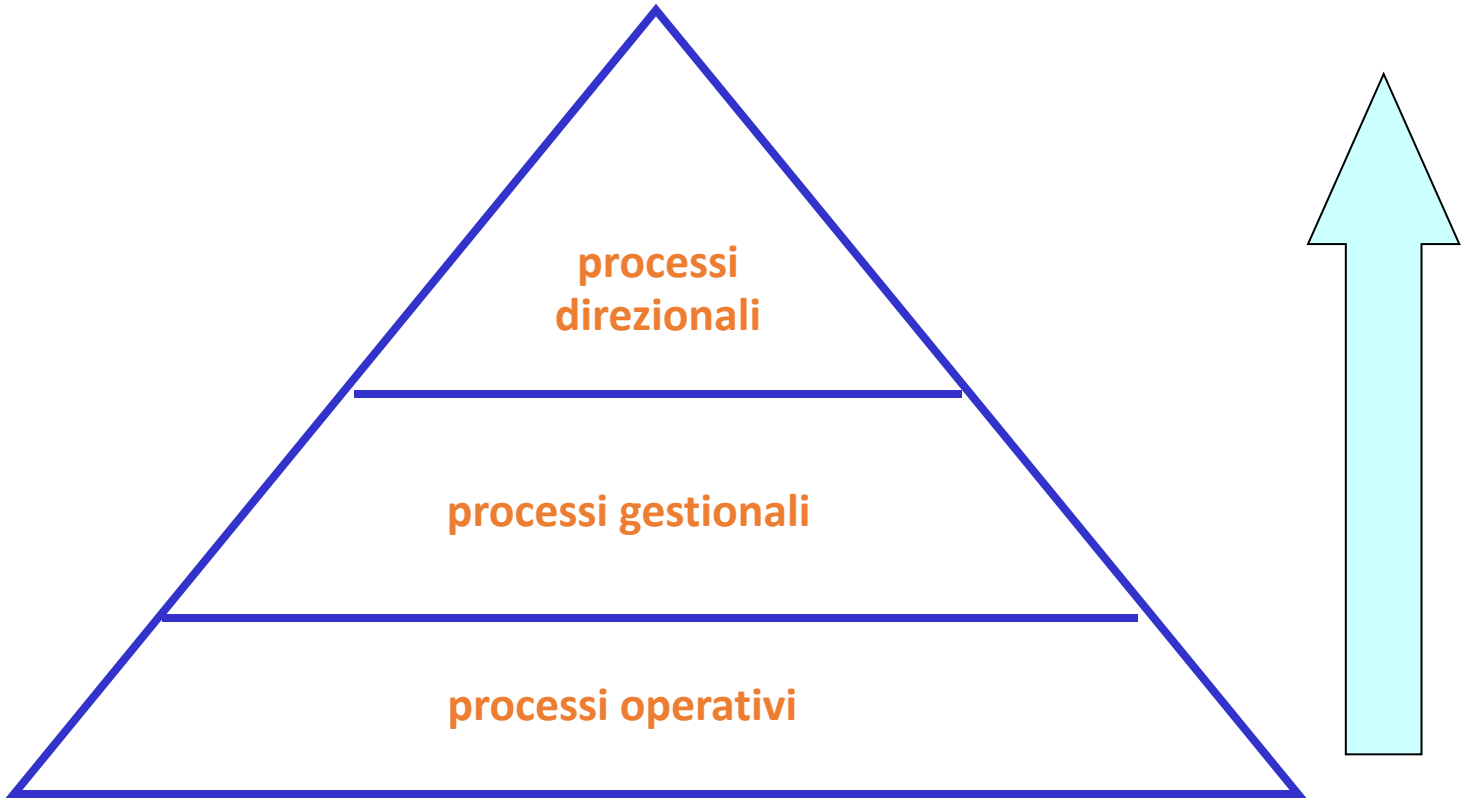
  o meglio, per il supporto ad essi
  - *Decision support systems* (il più possibile integrati)

Sistemi di supporto alle decisioni

- La tecnologia utilizzata per rendere disponibili alla dirigenza aziendale elementi quantitativi utili per prendere decisioni tattico-strategiche in modo efficace e veloce

- Ma su quali dati?
  - quelli accumulati per i processi operativi e gestionali

Processi e dati



Esigenze diverse: OLTP e OLAP

- nei sistemi di livello operativo
  - OLTP: On-Line Transaction Processing
- nei sistemi di livello più alto
  - OLAP: On-Line Analytical Processing

OLTP

- Tradizionale elaborazione di transazioni, che realizzano i processi operativi dell’azienda-ente
  - Operazioni
    - predefinite, brevi, (spesso) semplici
    - ogni operazione coinvolge “pochi” dati, nell'ambito di "un" processo
    - numerose
  - Dati di dettaglio, aggiornati
  - Le proprietà *ACID* (Atomicità, Consistenza, Isolamento, Durabilità) delle transazioni sono essenziali

OLAP

- Elaborazione di operazioni per il supporto alle decisioni
  - Operazioni
    - complesse e casuali
    - ogni operazione può coinvolgere molti dati, anche di processi diversi
  - Dati aggregati, storici, anche non attualissimi
  - Le proprietà ACID non sono rilevanti, perché le operazioni sono di sola lettura

OLTP e OLAP

	OLTP	OLAP
Utente	impiegato	dirigente
Funzione	operazioni giornaliere	supporto alle decisioni
Progettazione	orientata all'applicazione	orientata ai dati
Dati	correnti, aggiornati, dettagliati, relazionali, omogenei	storici, aggregati, multidimensionali, eterogenei
Uso	ripetitivo	casuale
Accesso	read-write, indicizzato	read, sequenziale
Unità di lavoro	transazione breve	interrogazione complessa
Record acc.	decine	milioni
N. utenti	migliaia	centinaia
Dimensione	100MB - 1GB	100GB - 1TB
Metrica	throughput	tempo di risposta

OLTP e OLAP

- I requisiti sono quindi contrastanti
- Le applicazioni dei due tipi possono danneggiarsi a vicenda

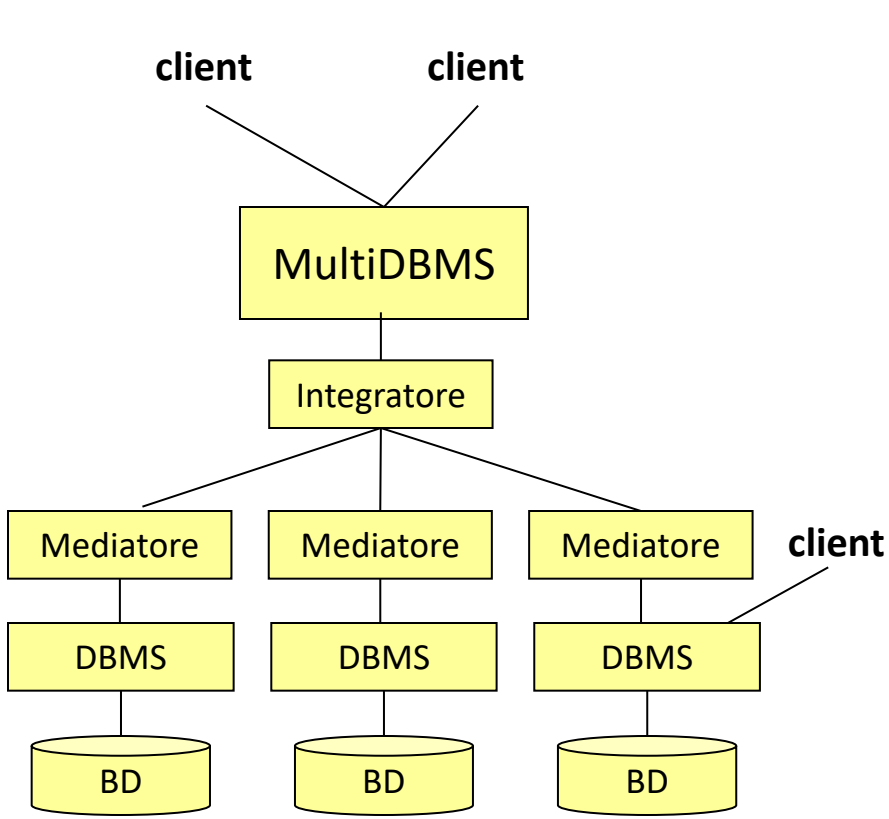
Evoluzione dei DSS (idea schematica)

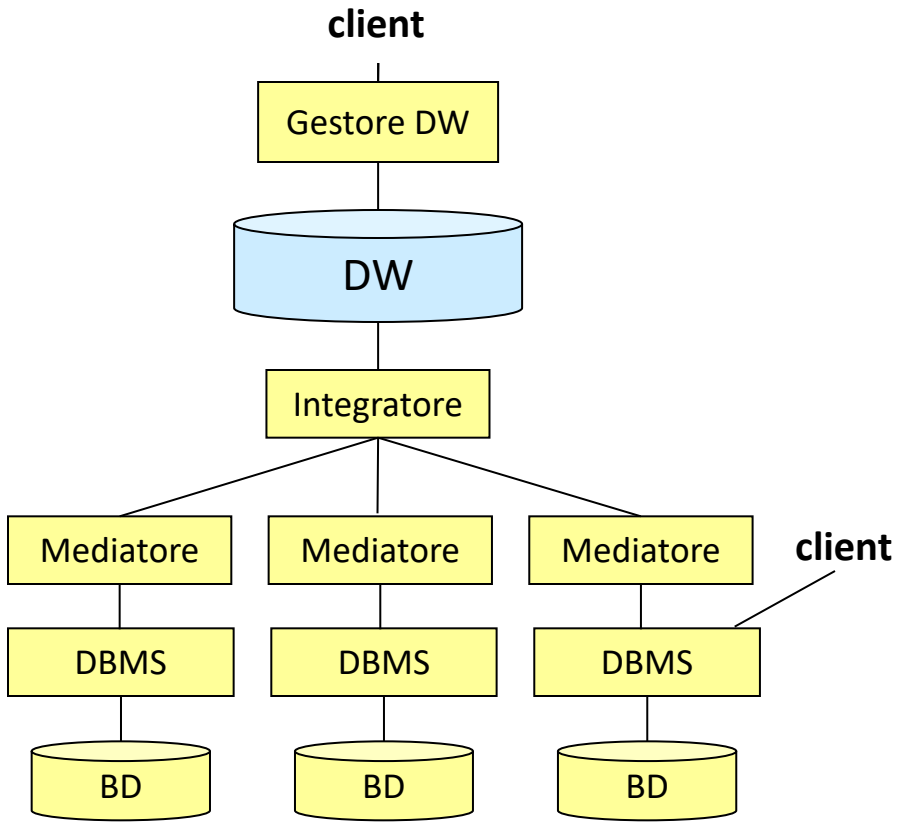
- Anni ‘60 — rapporti batch
  - difficile trovare e analizzare dati
  - ogni richiesta richiede un nuovo programma
- Anni ‘70 — DSS basato su terminale
  - accesso ai dati operazionali, molto inefficiente
- Anni ‘80 — strumenti d’automazione d’ufficio e di analisi
  - fogli elettronici, interfacce grafiche
- Anni ‘90 — data warehousing
  - strumenti di OLAP

L’obiettivo ideale è sensato e praticabile?

- La realtà è in continua evoluzione, non esiste uno “stato stazionario” (se non nell’iperuranio):
  - cambiano le esigenze
  - cambiano le strutture
  - le realizzazioni richiedono tempo
- Il coordinamento forte fra i vari settori può risultare controproducente
- Ogni organizzazione ha di solito diverse basi di dati *distribuite, eterogenee, autonome*

Multi-database e Data Warehouse (due approcci all'integrazione)





Data warehouse

Una base di dati

- utilizzata principalmente per il supporto alle decisioni direzionali o anche a livello più basso (OLAP e non OLTP)
- integrata — aziendale e non dipartimentale
- orientata ai dati — non alle applicazioni
- con dati storici — con un ampio orizzonte temporale, e indicazione (di solito) di elementi di tempo
- con dati aggregati (di solito) — per effettuare stime e valutazioni
- fuori linea — i dati sono aggiornati periodicamente
- separata dalle basi di dati operazionali

… integrata …

- I dati di interesse provengono da tutte le sorgenti informative — ciascun dato proviene da una o più di esse

- Il data warehouse rappresenta i dati in modo univoco — riconciliando le eterogeneità dalle diverse rappresentazioni
  - nomi
  - struttura
  - codifica
  - rappresentazione multipla

… orientata ai dati …

- Le basi di dati operazionali sono costruite a supporto dei singoli processi operativi o applicazioni
  - produzione
  - vendita
- Il data warehouse è costruito attorno alle principali entità del patrimonio informativo aziendale
  - prodotto
  - cliente

… dati storici …

- Le basi di dati operazionali mantengono il valore corrente delle informazioni
  - L'orizzonte temporale di interesse è dell'ordine dei pochi mesi

- Nel data warehouse è di interesse l'evoluzione storica delle informazioni
  - L'orizzonte temporale di interesse è dell'ordine degli anni

… dati aggregati …

- Nelle attività di analisi dei dati per il supporto alle decisioni
  - non interessa “chi” ma “quanti”
  - non interessa un dato ma
    - la somma,
    - la media,
    - il minimo e il massimo, …

  di un insieme di dati.

- Le operazioni di *aggregazione* sono quindi fondamentali nel data warehousing e nella costruzione/mantenimento di un data warehouse.

… fuori linea …

- In una base di dati operazionale, i dati vengono
  - acceduti
  - inseriti
  - modificati
  - cancellati

  pochi record alla volta

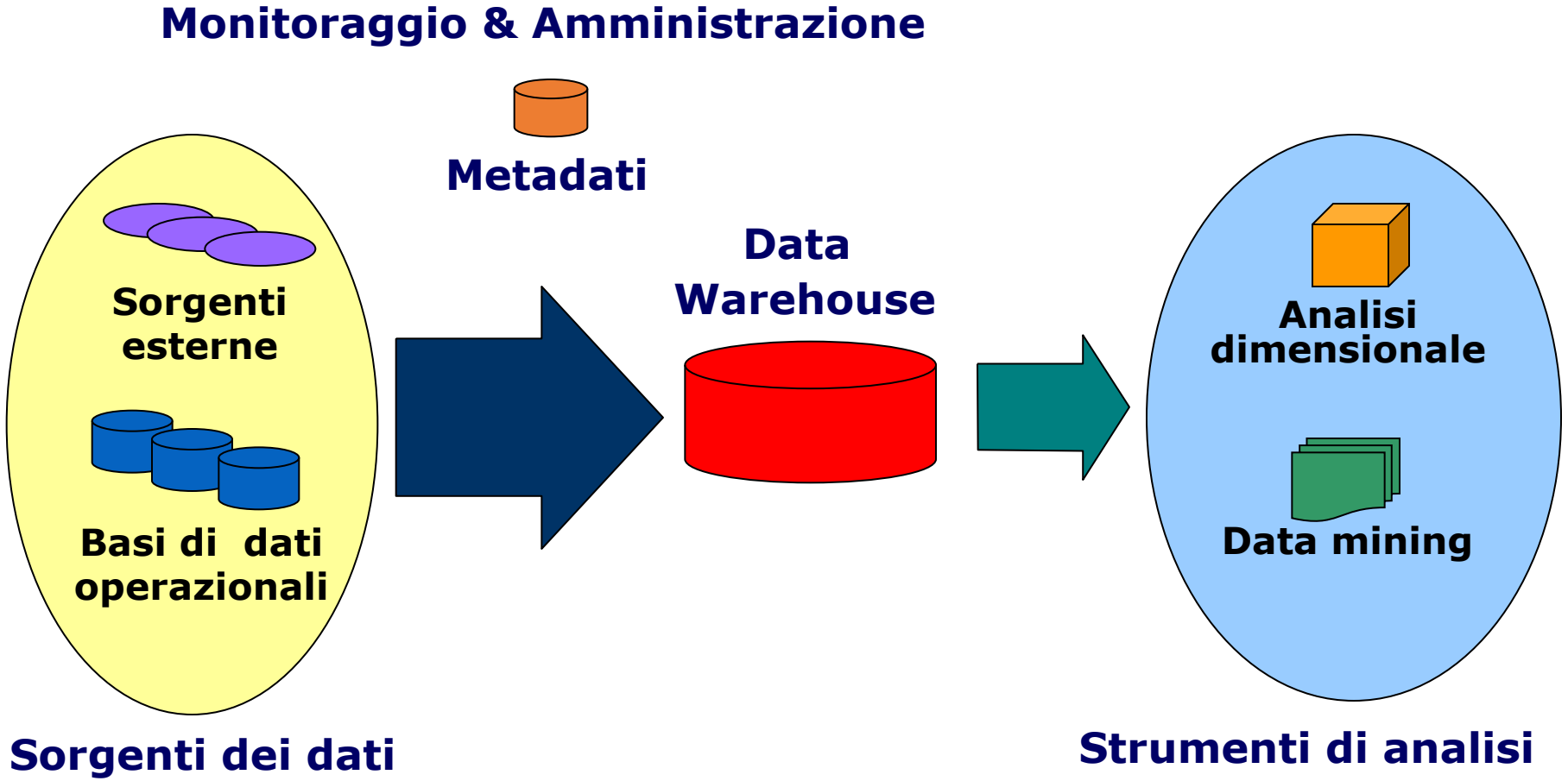
- Nel data warehouse, abbiamo
  - operazioni di accesso e interrogazione — “diurne”
  - operazioni di caricamento e aggiornamento dei dati — “notturne”

  che riguardano milioni di record

… una base di dati separata …

- Un data warehouse viene mantenuto separatamente dalle basi di dati operazionali perché
  - non esiste un'unica base di dati operazionale che contiene tutti i dati di interesse
  - la base di dati deve essere integrata
  - non è tecnicamente possibile fare l'integrazione in linea; degrado generale delle prestazioni senza la separazione
    - l'analisi dei dati richiede per i dati organizzazioni speciali e metodi di accesso specifici
  - i dati di interesse sarebbero comunque diversi
    - devono essere mantenuti dati storici
    - devono essere mantenuti dati aggregati

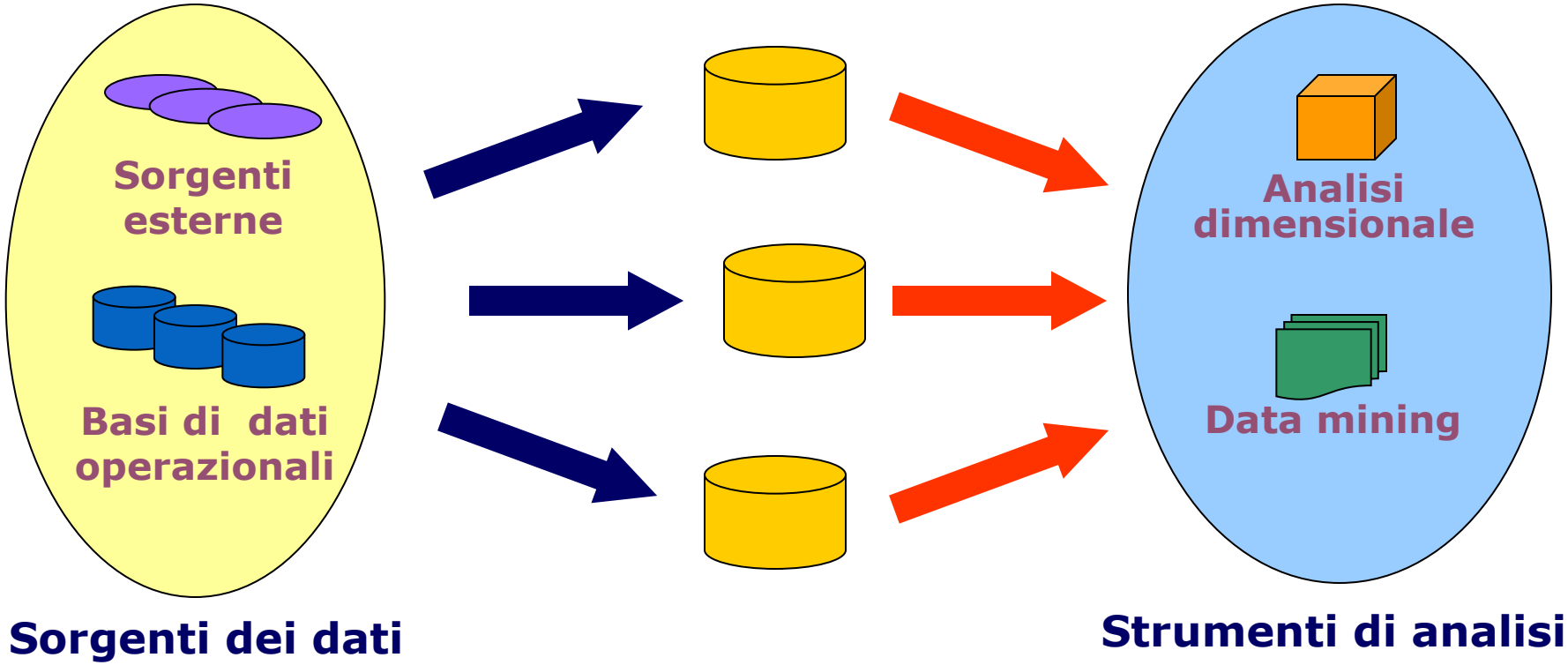
Architettura per il data warehousing



Esigenze di analisi e integrazione

- Molto spesso:
  - l'analisi è mirata a specifici processi della azienda o ente
  - un vero e proprio DW integrato
    - non interessa
    - non "viene in mente"
    - non si riesce a fare (per urgenza, mancanza di risorse, o mancanza di "competenza e responsabilità")
  - può essere utile o necessario concentrarsi (almeno temporaneamente) su un suo sottoinsieme

Architettura “realistica”

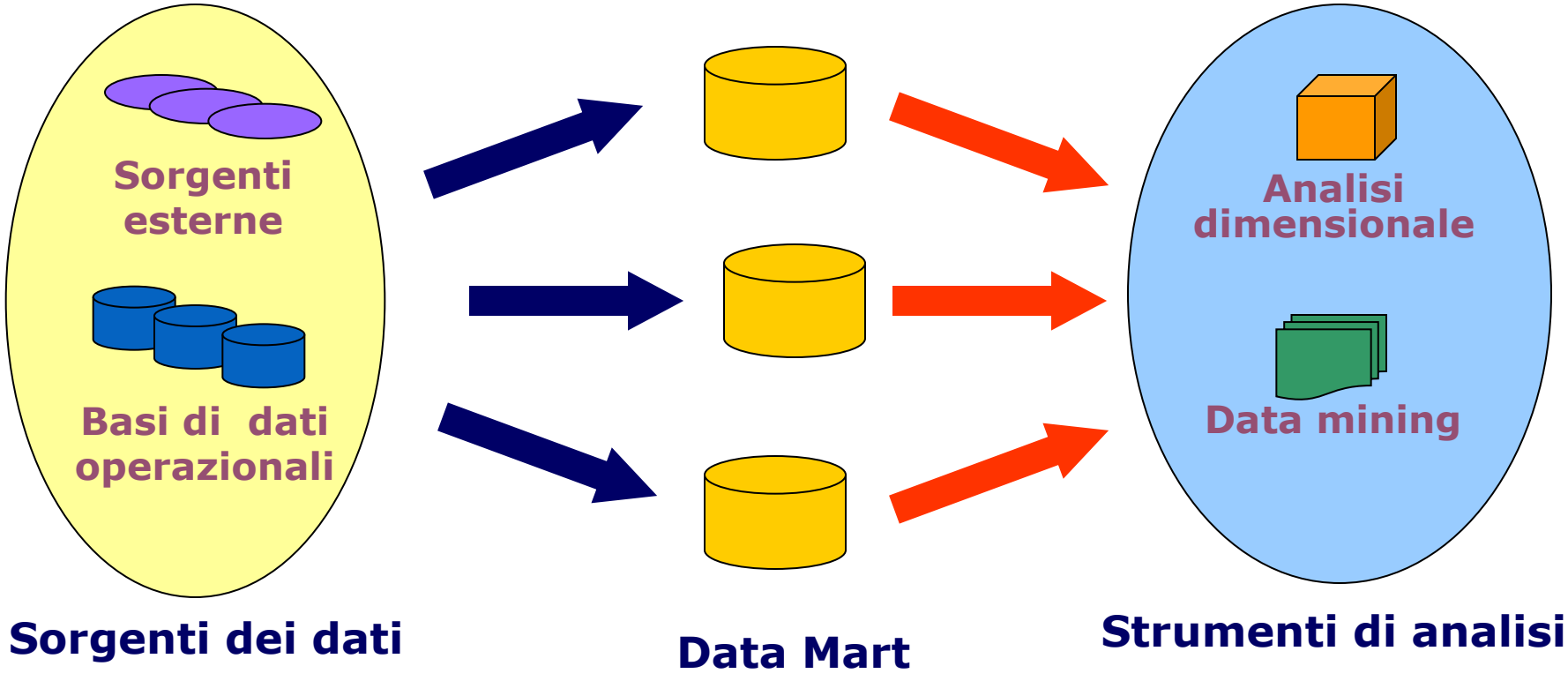


Data mart

- Un sottoinsieme logico dell'intero data warehouse
  - un data mart è la restrizione del data warehouse a un singolo processo
  - un data warehouse è l'unione di tutti i suoi data mart

    (il che non è detto che vada sempre bene, vediamo fra poco)

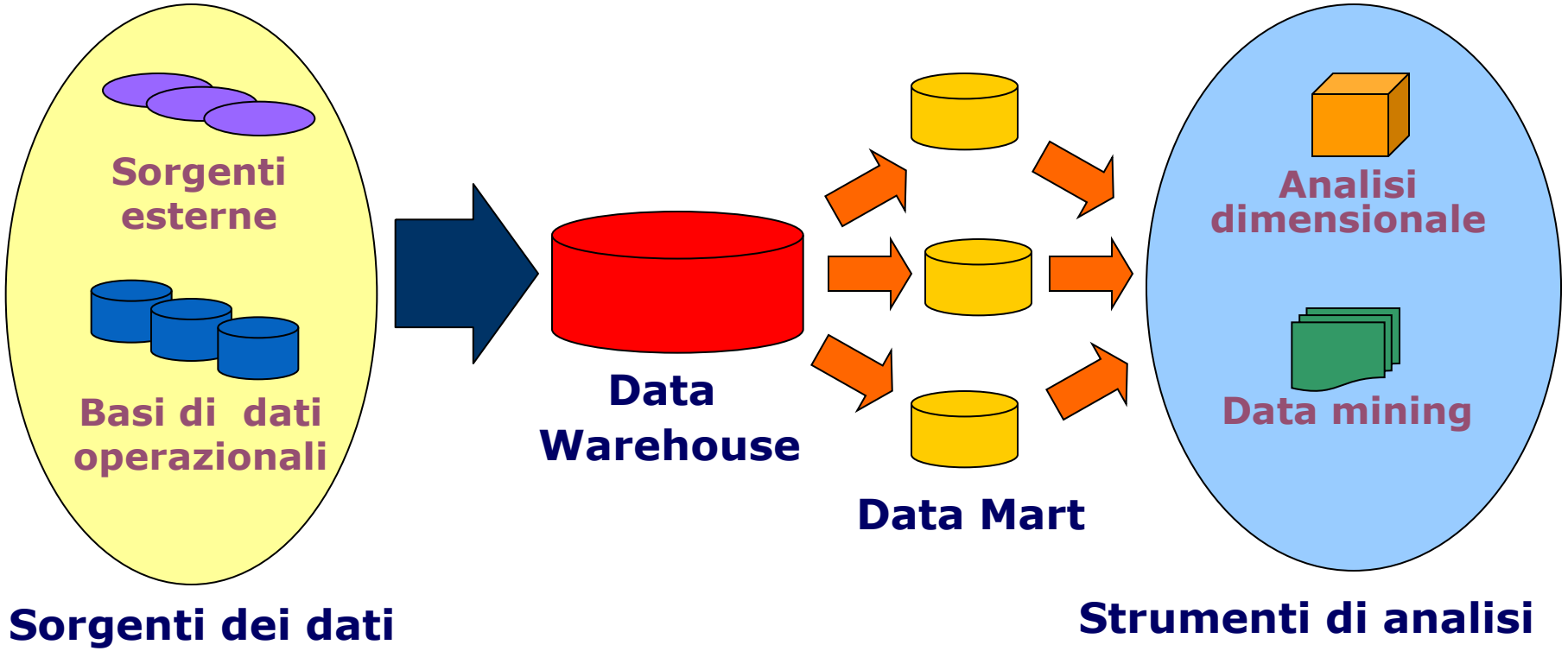
Architettura “realistica”



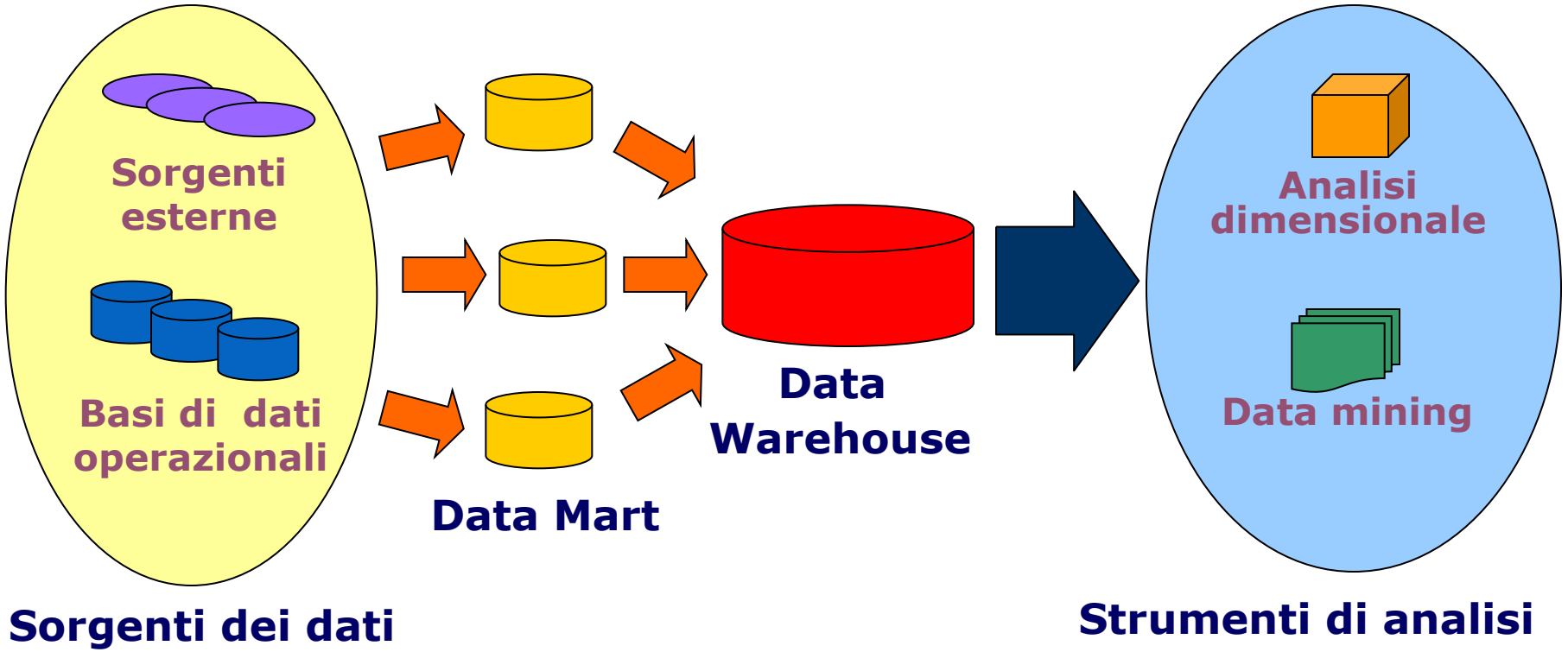
Top-down o bottom-up?

- Prima il data warehouse o prima i data mart?

DW e DM


Sorgenti esterne
Basi di dati operazionali
Data Warehouse
Data Mart
Analisi dimensionale
Data mining
Sorgenti dei dati
Strumenti di analisi

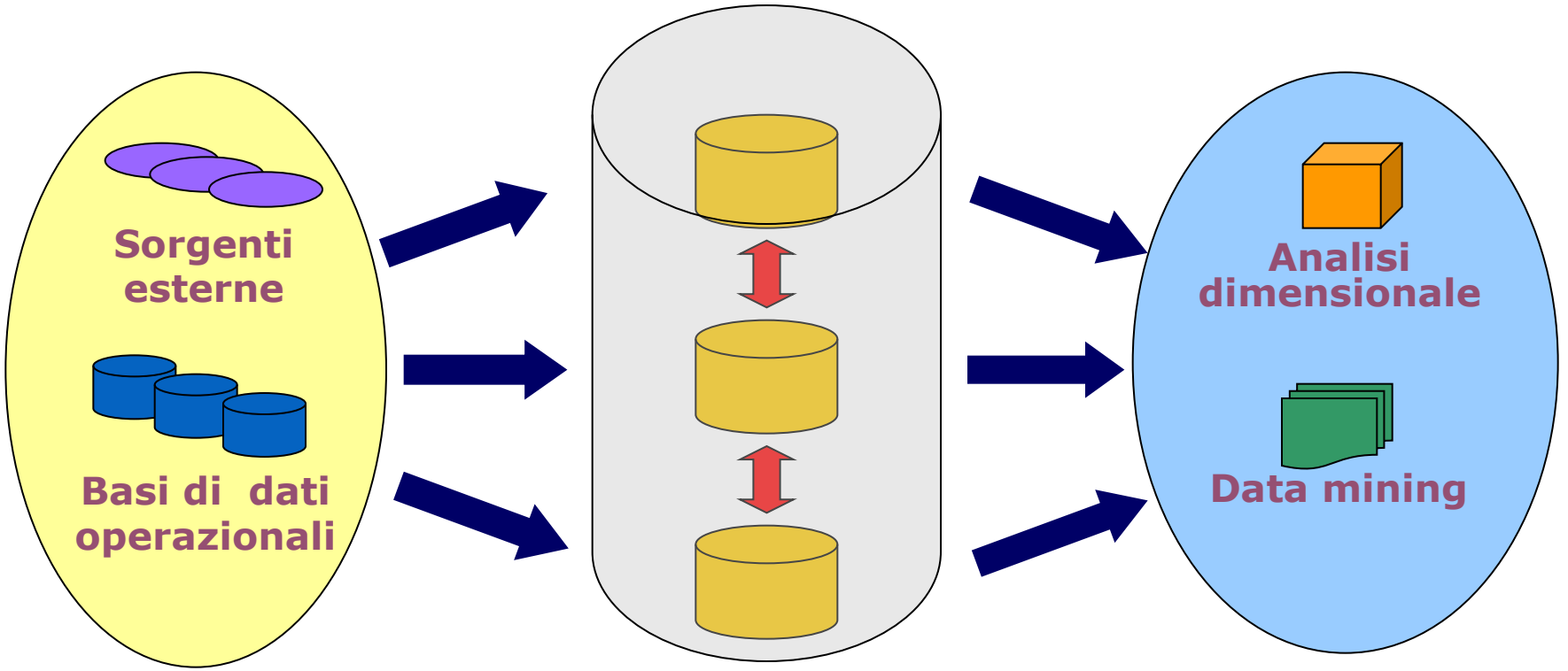
DW e DM


Sorgenti esterne
Basi di dati operazionali
Data Mart
Data Warehouse
Analisi dimensionale
Data mining
Sorgenti dei dati
Strumenti di analisi

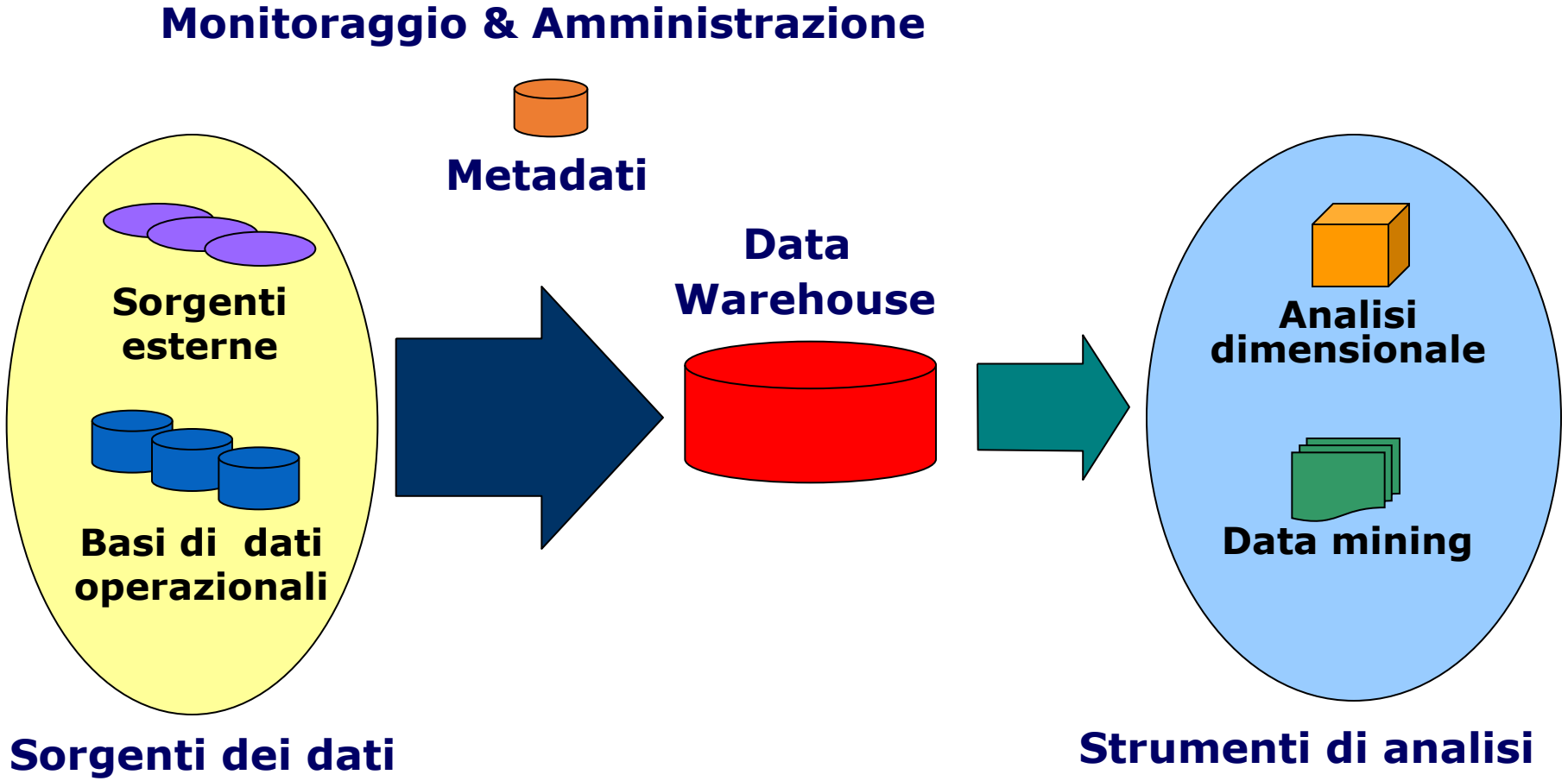
Data mart e DW

- Prima il data warehouse o prima i 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 di un “buon” data warehouse
- Non c'è risposta, o meglio: nessuno dei due!
- Infatti:
  - l'approccio è spesso incrementale
- Ma
  - è necessario coordinare i data mart:
    - dimensioni conformi e “DW bus”

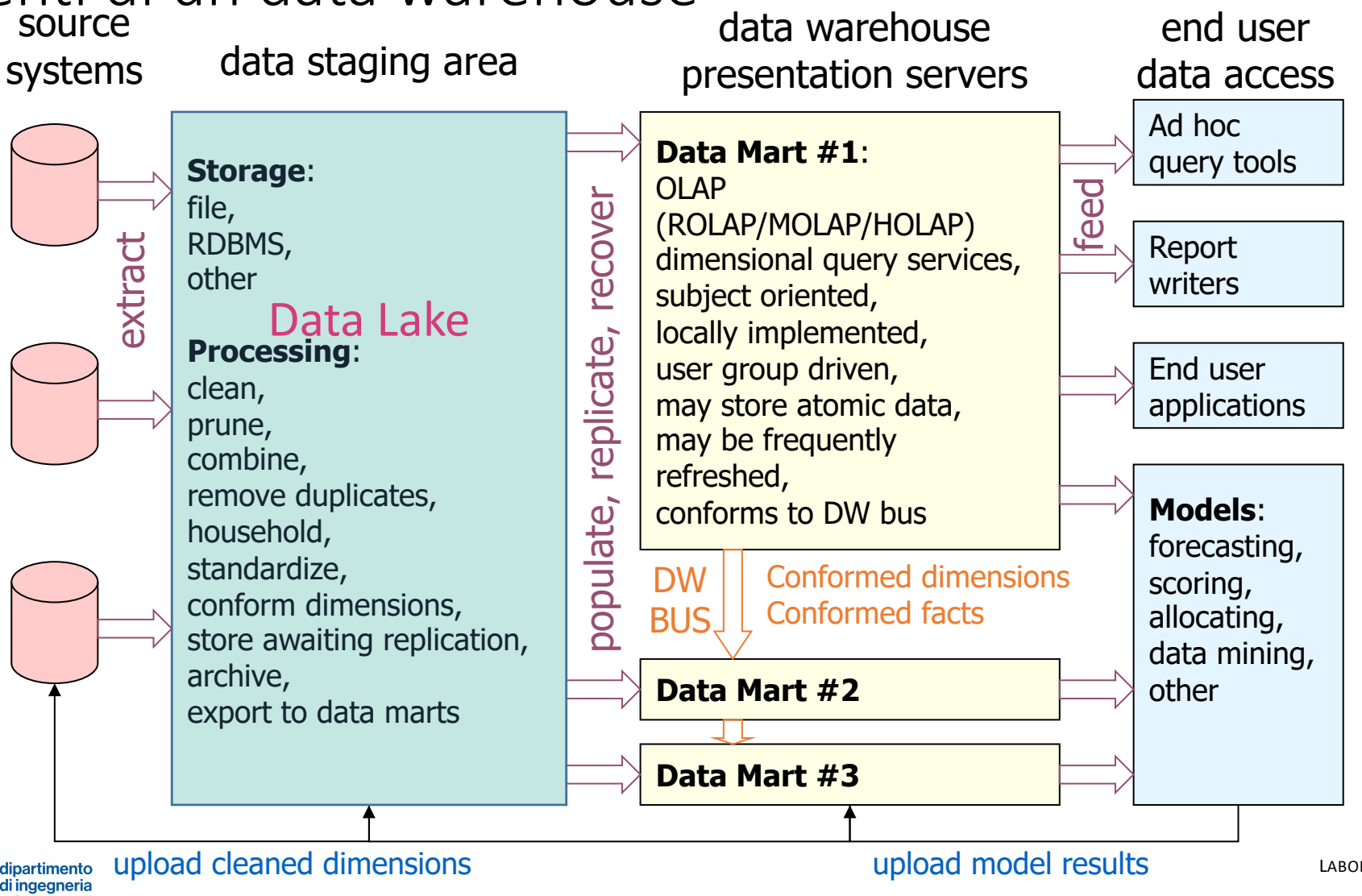
DM e DW


Sorgenti esterne
Basi di dati operazionali
Analisi dimensionale
Data mining
Sorgenti dei dati
Data Mart
DW con "bus"
Strumenti di analisi

Architettura per il data warehousing



Elementi di un data warehouse



Sorgenti informative

- 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
    - *non in modo conforme nell'ambito dell'organizzazione*
  - sono sistemi "legacy"
- sorgenti esterne
  - ad esempio, dati forniti da società specializzate di analisi

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
  - richiede un insieme complesso di attività semplici
  - è distribuita su più calcolatori e ambienti eterogenei
  - gestisce i dati prevalentemente con formati di varia natura (spesso semplici file)

ETL

- Extract, Transform, Load
- Il processo (complesso) che porta i dati dai sistemi operazionali al data warehouse, passando per l'area di staging

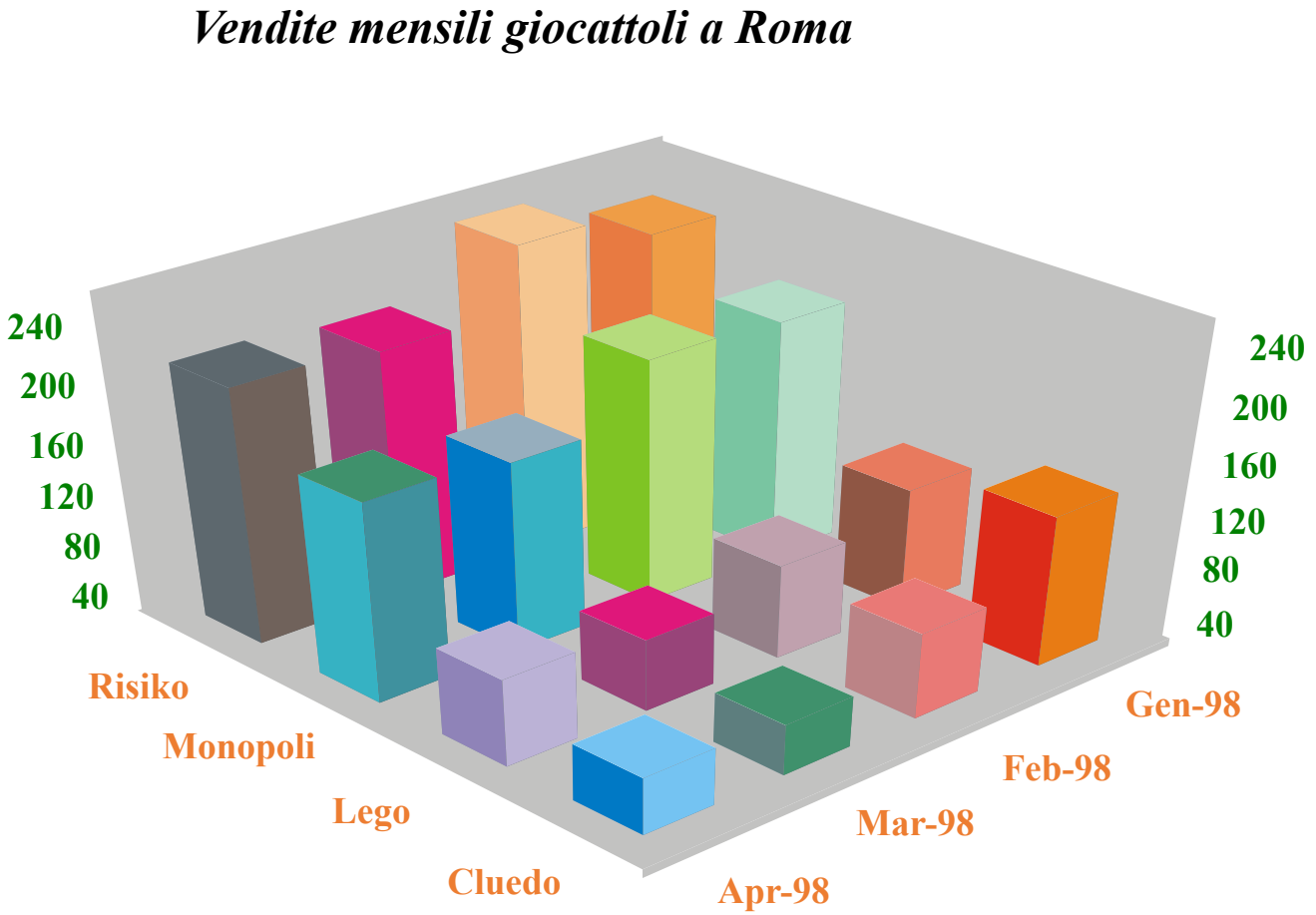
Server di presentazione

- 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 **multidimensionale**
    (secondo i concetti di fatto e dimensione, vediamo fra poco)
  - tecnologie che possono essere adottate
    - RDBMS: ROLAP
    - tecnologia OLAP esplicita: MOLAP
      - i concetti di fatto e dimensione sono espliciti

Visualizzazione dei dati

- I dati vengono infine visualizzati in veste grafica, in maniera da essere facilmente comprensibili.
- Si fa uso di:
  - tabelle
  - istogrammi
  - grafici
  - torte
  - superfici 3D
  - bolle
  - area in pila
  - forme varie
  - …

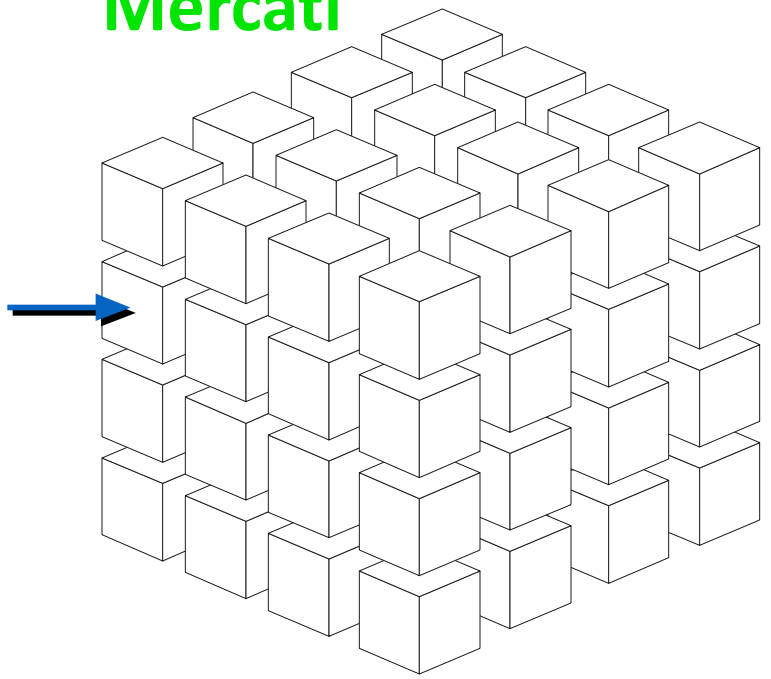
Visualizzazione finale di un’analisi



Rappresentazione multidimensionale

VENDITE

Mercati



Incassi

Prodotti

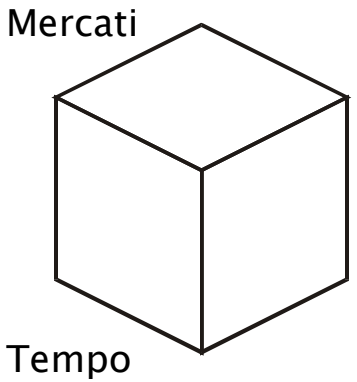
Periodi di tempo

anche Costi, Guadagni

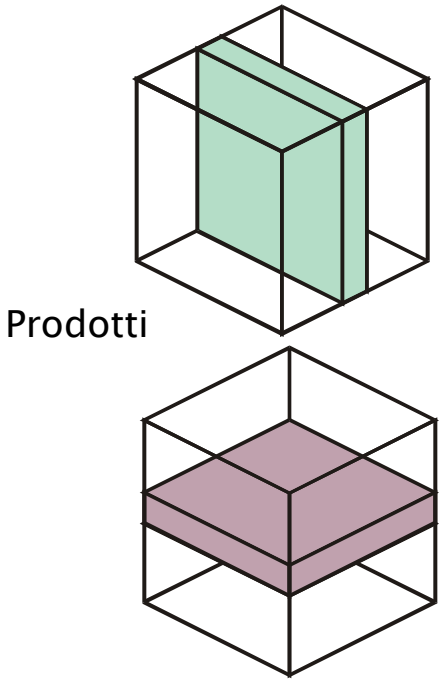
Modello “logico” per DW

- L’analisi dei dati avviene rappresentando i dati in forma *multidimensionale*
- Concetti rilevanti:
  - *fatto* — un concetto sul quale centrare l’analisi
  - *misura* — una proprietà atomica di un fatto da analizzare
  - *dimensione* — descrive una prospettiva lungo la quale effettuare l’analisi
- Esempi di fatti/misure/dimensioni
  - vendita / quantità venduta, incasso / prodotto, tempo
  - telefonata / costo, durata / chiamante, chiamato, tempo

Viste su dati multidimensionali

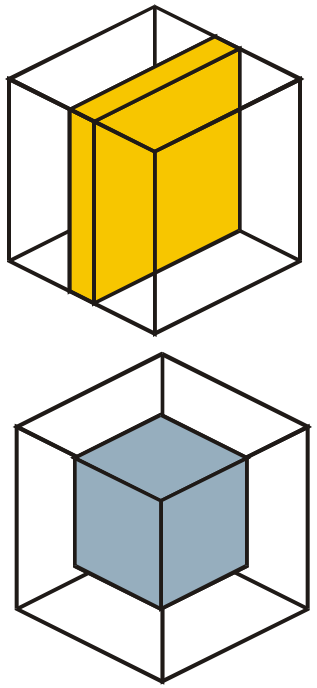


Il manager regionale esamina la vendita dei prodotti in tutti i periodi relativamente ai propri mercati



Il manager di prodotto esamina la vendita di un prodotto in tutti i periodi e in tutti i mercati

Il manager finanziario esamina la vendita dei prodotti in tutti i mercati relativamente al periodo corrente e quello precedente



Il manager strategico si concentra su una categoria di prodotti, una area e un orizzonte temporale

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
	90	26	53	32	32	48

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

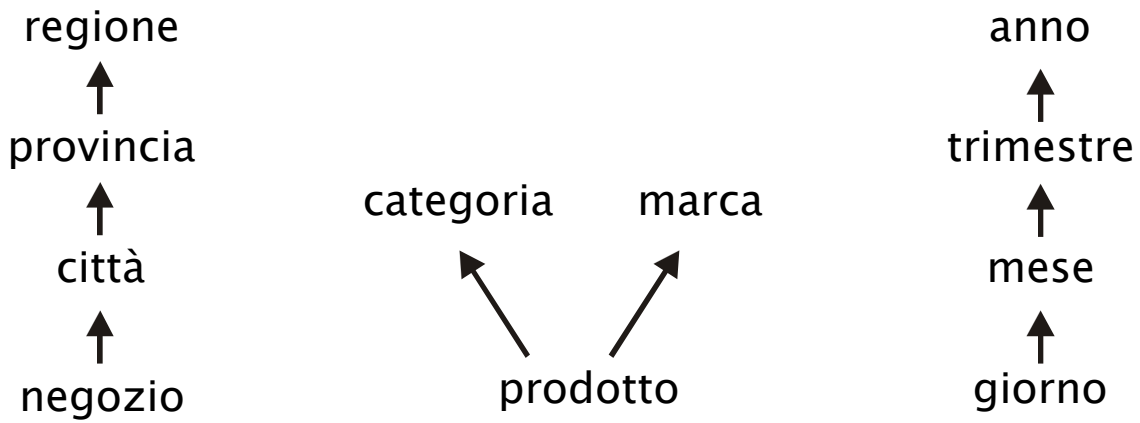
Pisa	38
Firenze 1	52
Firenze 2	27
Roma 1	56
Roma 2	34
Roma 3	56
Latina	18

Operazioni su dati multidimensionali

- *Roll up* (o *drill up*) — aggrega i dati
  - volume di vendita totale dello scorso anno per categoria di prodotto e regione
- *Drill down* — disaggrega i dati
  - per una particolare categoria di prodotto e regione, mostra le vendite giornaliere dettagliate per ciascun negozio
- *Slice & dice* — seleziona e proietta
- *Pivot* — re-orienta il cubo

Dimensioni e gerarchie di livelli

- Ciascuna dimensione è organizzata in una gerarchia che rappresenta i possibili livelli di aggregazione per i dati
  - negozio, città, provincia, regione
  - prodotto, categoria, marca
  - giorno, mese, trimestre, anno



	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze	25	8	14	10	12	10
Roma	50	13	24	18	12	29
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Toscana	37	10	24	13	18	15
Lazio	53	16	29	19	14	33

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	I trim	II trim
Pisa	24	14
Firenze 1	35	17
Firenze 2	12	15
Roma 1	28	28
Roma 2	23	11
Roma 3	36	20
Latina	11	7

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	I trim	II trim
Pisa	24	14
Firenze 1	35	17
Firenze 2	12	15
Roma 1	28	28
Roma 2	23	11
Roma 3	36	20
Latina	11	7

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze	25	8	14	10	12	10
Roma	50	13	24	18	12	29
Latina	3	3	5	1	2	4

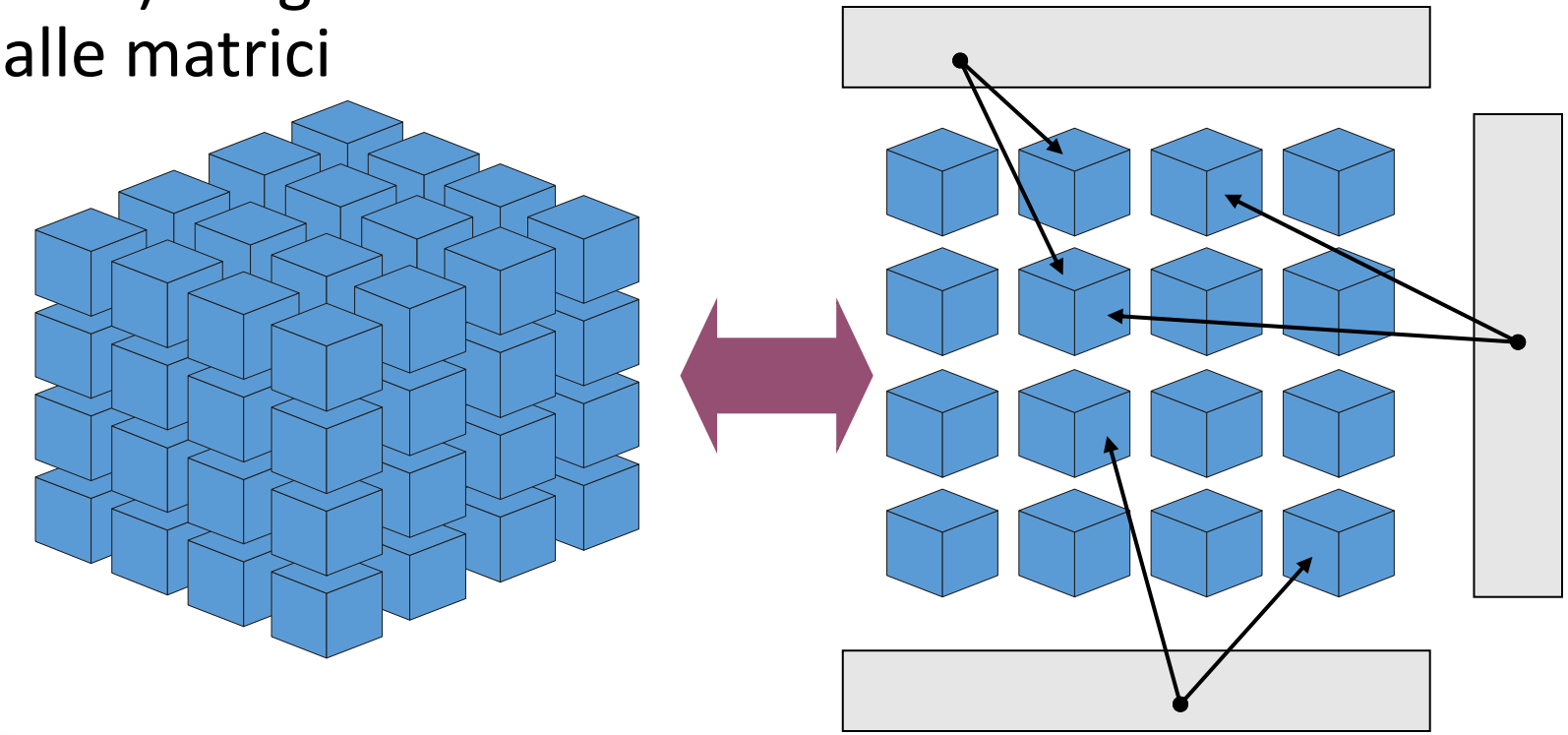
	I trim	II trim
Pisa	24	14
Firenze	47	32
Roma	87	59
Latina	11	7

Implementazione per dati multidimensionali

- MOLAP
  - M = multidimensional
- ROLAP
  - R = relational

Implementazione MOLAP

- I dati sono memorizzati direttamente in un formato dimensionale (proprietario). Le gerarchie sui livelli sono codificate in indici di accesso alle matrici



Implementazione ROLAP: schemi dimensionali

- Uno *schema dimensionale* (*schema a stella, star schema*) è composto da
  - una tabella principale, *tabella fatti*
    - la tabella fatti memorizza le misure di un processo
    - i fatti più comuni hanno misure numeriche e additive
  - due o più tabelle ausiliarie, *tabelle dimensione*
    - una tabella dimensione rappresenta una prospettiva, un aspetto rispetto a cui è interessante analizzare i fatti
    - gli attributi sono solitamente testuali, discreti e descrittivi
- Intuitivamente:
  - rappresentazione sparsa di una matrice multidimensionale
  - relationship n-aria

Schema dimensionale

CodNegozio	Nome
PI	Pisa
FI1	Firenze 1
FI2	Firenze 2
RM1	Roma 1
RM2	Roma 2
RM3	Roma 3
LT	Latina

CodNegozio	CodMese	Vendite
PI	Gen	12
PI	Feb	2
PI	Mar	10
PI	Apr	3
PI	Mag	6
PI	Giu	5
FI1	Gen	21
FI1	Feb	4
FI1	Mar	10
FI1	Apr	4
FI1	Mag	6
FI1	Giu	7
…	…	…

CodMese	Mese
Gen	gennaio
Feb	febbraio
Mar	marzo
Apr	aprile
Mag	maggio
Giu	giugno

Schema dimensionale

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

CodNegozio	Nome
PI	Pisa
FI1	Firenze 1
FI2	Firenze 2
RM1	Roma 1
RM2	Roma 2
RM3	Roma 3
LT	Latina

CodNegozio	CodMese	Vendite
PI	Gen	12
PI	Feb	2
PI	Mar	10
PI	Apr	3
PI	Mag	6
PI	Giu	5
FI1	Gen	21
FI1	Feb	4
FI1	Mar	10
FI1	Apr	4
FI1	Mag	6
FI1	Giu	7
…	…	…

CodMese	Mese
Gen	gennaio
Feb	febbraio
Mar	marzo
Apr	aprile
Mag	maggio
Giu	giugno

Schema dimensionale: dimensioni con livelli

CodN	…	Città	Regione	…
PI	…	Pisa	Toscana	…
FI1	…	Firenze	Toscana	…
FI2	…	Firenze	Toscana	…
RM1	…	Roma	Lazio	…
RM2	…	Roma	Lazio	…
RM3	…	Roma	Lazio	…
LT	…	Latina	Lazio	…

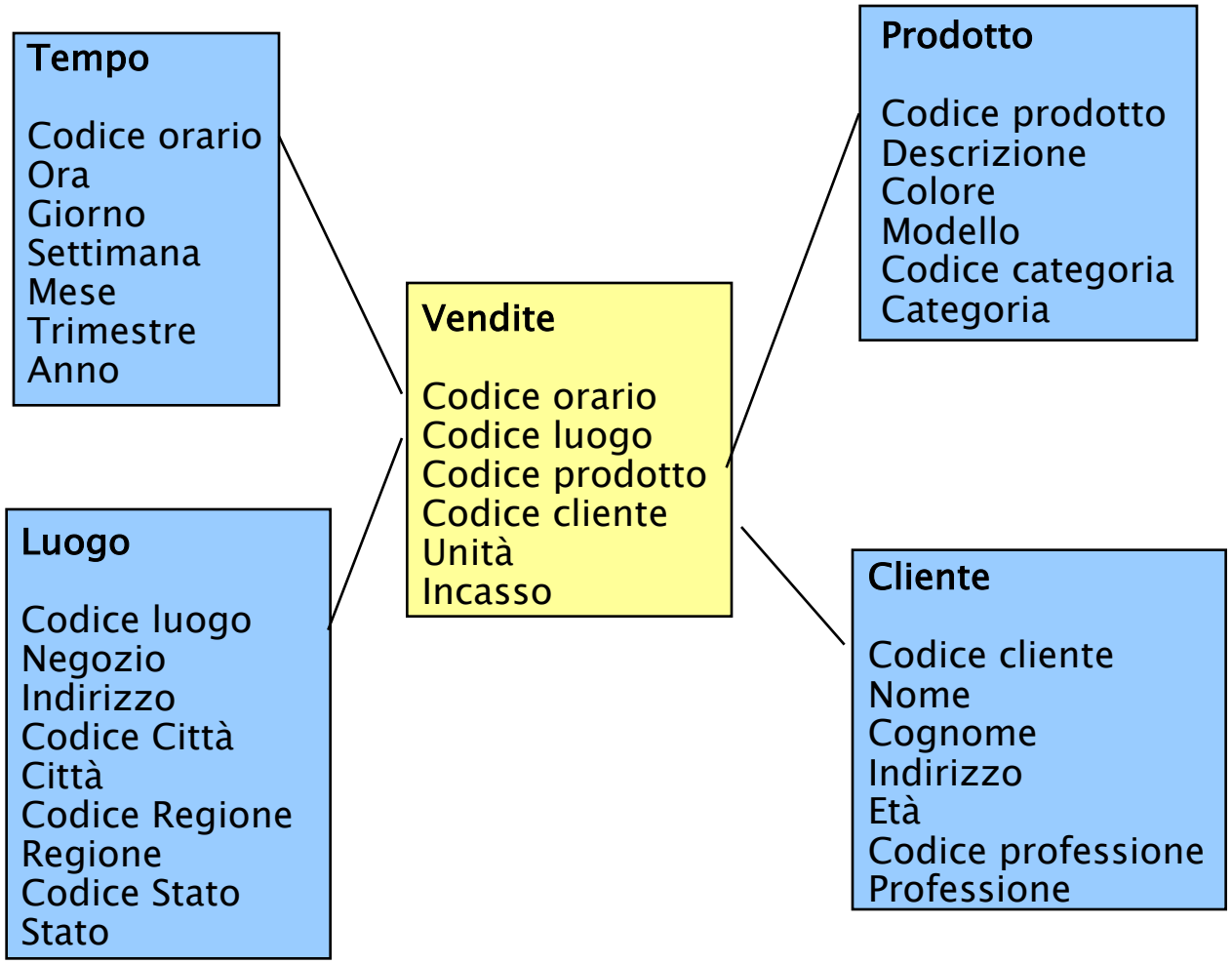
CodN	CodM	Vendite
PI	Gen	12
PI	Feb	2
PI	Mar	10
PI	Apr	3
PI	Mag	6
PI	Giu	5
FI1	Gen	21
FI1	Feb	4
FI1	Mar	10
FI1	Apr	4
FI1	Mag	6
FI1	Giu	7
…	…	…

CodM	Mese	Trimestre
Gen	gennaio	I trim
Feb	febbraio	I trim
Mar	marzo	I trim
Apr	aprile	II trim
Mag	maggio	II trim
Giu	giugno	II trim

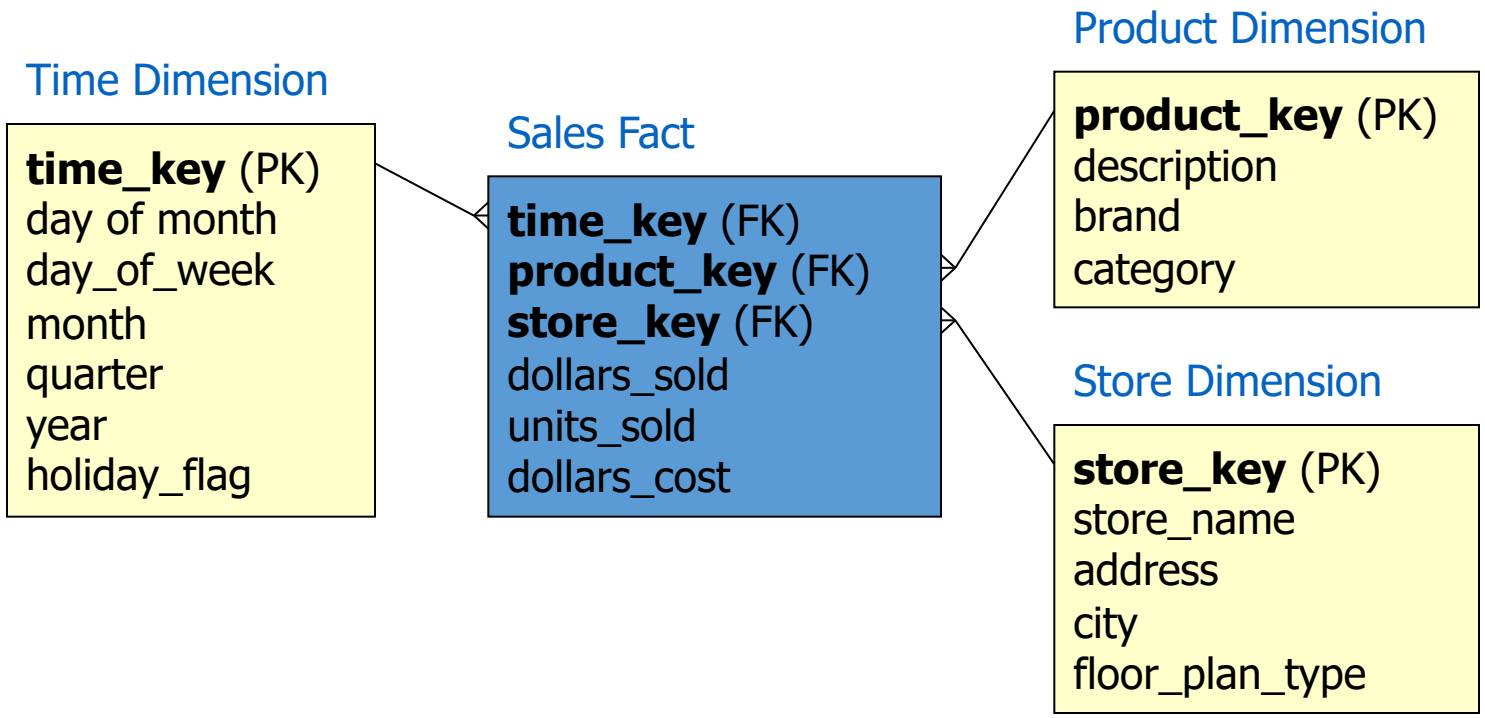
Data warehouse dimensionale

- lo schema di un data warehouse è un insieme di schemi dimensionali
  - ogni data mart è un insieme di schemi dimensionali
  - tutti i data mart vengono costruiti usando il “DW bus”
    - dimensioni conformi
    - ogni dimensione ha lo stesso significato in ciascuno schema dimensionale e data mart
    - le ennuple sono le stesse (o comunque in rapporto uno a uno; potrebbero essere sottoinsiemi, ma allora ne deve esistere una versione "completa")
    - fatti conformi
    - anche i fatti hanno interpretazione uniforme

Uno schema dimensionale



Un altro schema dimensionale



- i dati delle vendite di prodotti in un certo numero di negozi nel corso del tempo
  - memorizza i totali delle vendite di un certo prodotto in un certo giorno in un certo negozio

Schemi dimensionali, dettagli

- Dimensioni
  - tabelle dimensione, caratteristiche
  - chiavi
  - "snowflaking"
- Fatti
  - tabelle fatti, caratteristiche
  - additività

Tabella dimensione

- Memorizza gli elementi (o membri) di una dimensione rispetto alla quale è interessante analizzare un processo (e le relative descrizioni)
- Ciascun record di una tabella dimensione descrive esattamente un elemento della rispettiva dimensione
  - un record di Time Dimension descrive un giorno (nell'ambito dell'intervallo temporale di interesse), in quanto il giorno è il dettaglio (massimo) che interessa
  - un record di Product Dimension descrive un prodotto in vendita nei negozi

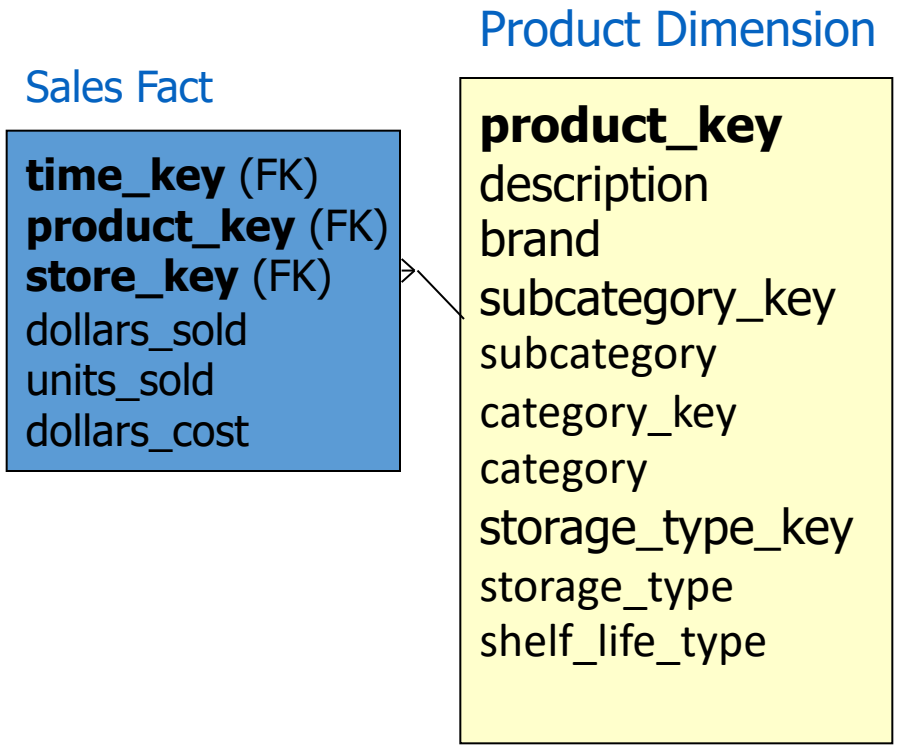
- I campi (non chiave) memorizzano gli attributi dei membri
  - gli attributi sono le proprietà dei membri, che sono solitamente testuali, discrete e descrittive

Chiavi nei DW

- Negli schemi dimensionali, si preferiscono di solito chiavi semplici (numeriche) e “locali” (progressive), per vari motivi
  - sono piccole (e evitano le chiavi composte)
  - permettono di gestire casi speciali (ad esempio, la “non appartenenza” ad una categoria)
  - evitano problemi dovuti al riuso (esempio, le matricole dei laureati, oppure le fatture che ricominciano da 1 ogni anno) o quelli dovuti alle fusioni aziendali
  - evitano i cambi di tipo (esempio, le targhe auto)

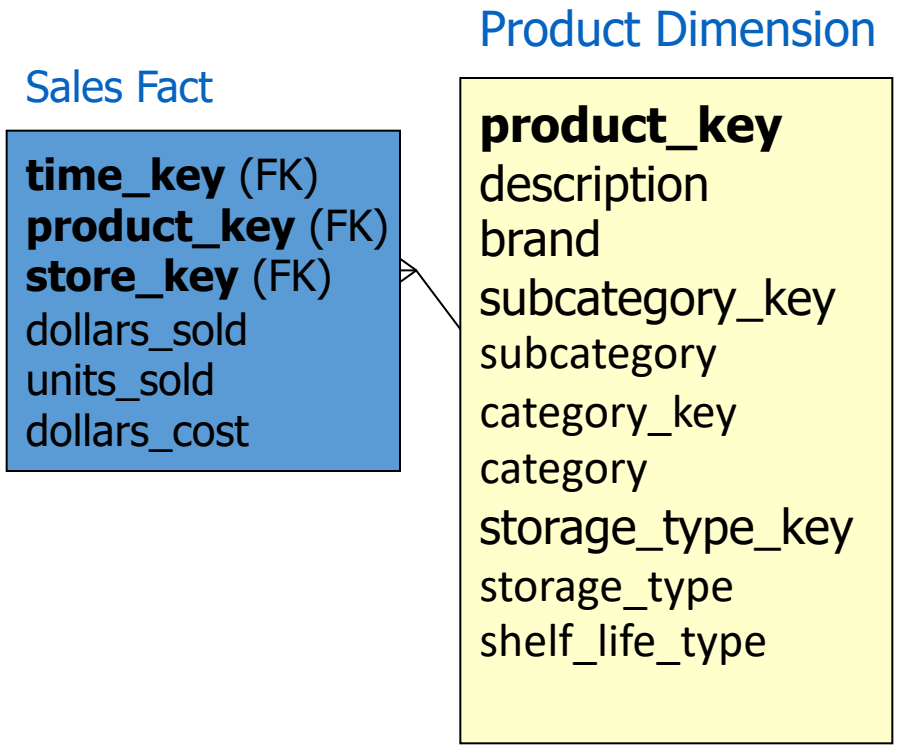
DW e normalizzazione

- Le dimensioni sono spesso *non normalizzate*



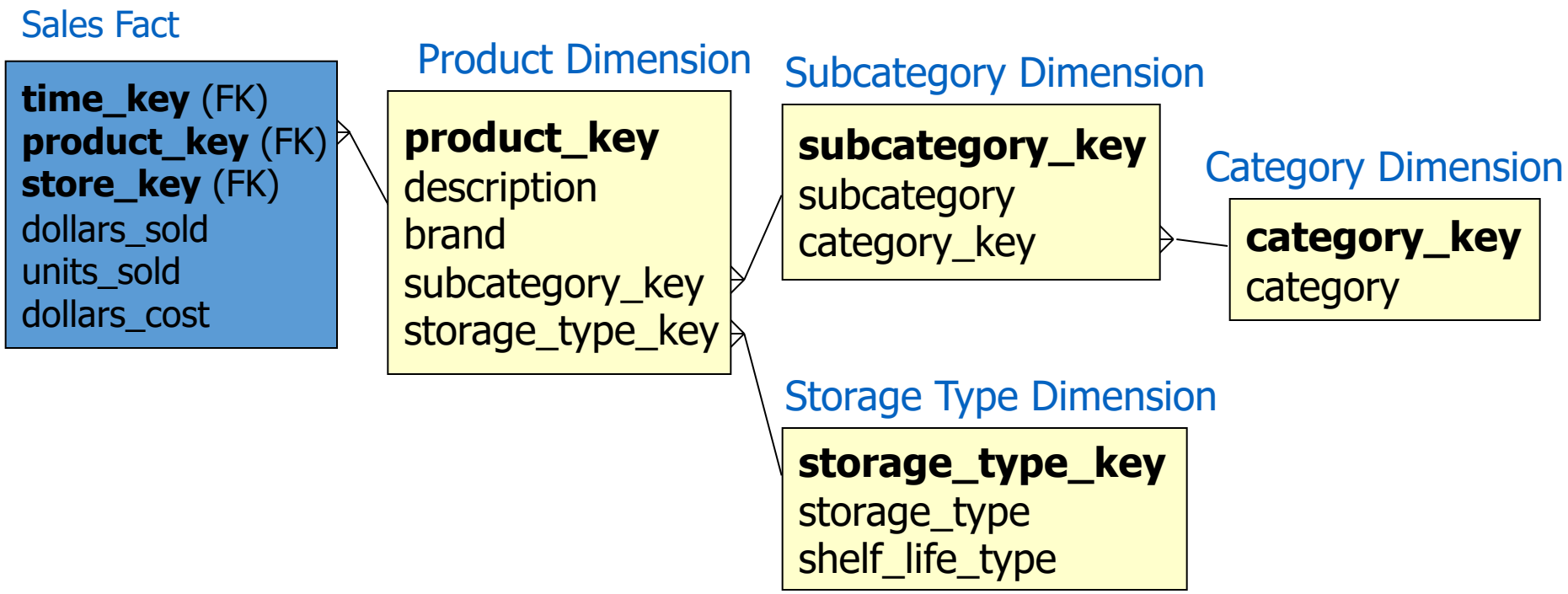
Snowflaking

- Normalizzazione di una tabella dimensione, che evidenzia “gerarchie di attributi”



Snowflaking

- Normalizzazione di una tabella dimensione, che evidenzia “gerarchie di attributi”



Occupazione di memoria

- Stima dell'occupazione di memoria della base di dati dimensionale di esempio
  - Tempo: 2 anni di 365 giorni, ovvero 730 giorni
  - Negozi: 300
  - Prodotti: 30.000
  - Fatti relativi alle vendite
    - ipotizziamo un livello di sparsità del 10% delle vendite giornaliere dei prodotti nei negozi ovvero, che ogni negozio vende giornalmente 3.000 diversi prodotti
    - 730 x 300 x 3000 = 630.000.000 record

Snowflaking: sintesi

- Lo snowflaking è solitamente svantaggioso
  - inutile per l'occupazione di memoria
    - ad esempio, supponiamo che la dimensione prodotto contenga 30.000 record, di circa 2.000 byte ciascuno occupando quindi 60MB di memoria
    - la tabella fatti contiene invece 630.000.000 record, di circa 10 byte ciascuno occupando quindi 6.3GB di memoria
    - le tabelle fatti sono sempre molto più grandi delle tabelle dimensione associate
      - anche riducendo l'occupazione di memoria della dimensione prodotto del 100%, l'occupazione di memoria complessiva è ridotta di meno dell'1%

Snowflaking: sintesi

- Lo snowflaking è solitamente svantaggioso
  - può peggiorare decisamente le prestazioni e rende più complessa e meno leggibile la scrittura delle interrogazioni
  - non porta a benefici in termini di riduzione di anomalie, perché le dimensioni non sono soggette ad aggiornamenti come nelle basi di dati transazionali

Tabella fatti

- Memorizza le misure numeriche di un processo
  - ogni record della tabella fatti memorizza una ennupla di misure (fatti) relativa a una combinazione degli elementi delle dimensioni ("all'intersezione di tutte le dimensioni") con riferimento alla granularità ("grana") scelta
- Nell'esempio
  - il processo (i fatti) è la vendita di prodotti nei negozi
  - le misure (i fatti) sono l'incasso in dollari (dollars_sold), la quantità venduta (units_sold),le spese sostenute a fronte della vendita (dollars_cost)
  - la grana è il totale per prodotto, negozio e giorno

Tabella fatti

- I campi della tabella fatti sono partizionati in due insiemi
  - chiave (composta)
    - sono *riferimenti alle chiavi primarie delle tabelle dimensione*
    - stabiliscono la grana della tabella fatti
  - altri campi: misure
    - talvolta chiamati proprio "fatti"
    - solitamente valori numerici comparabili e additivi (vediamo tra poco)
- Una tabella fatti memorizza una funzione (in senso matematico) dalle dimensioni ai fatti
  - ovvero, una funzione che associa (o meglio, può associare) un valore per ciascuna possibile combinazione dei membri delle dimensioni

Additività dei fatti

- Un fatto (o, meglio, una misura) è *additivo* se ha senso sommarlo (o *aggregarlo* in qualche modo) rispetto a ogni possibile combinazione delle dimensioni da cui dipende
  - l'incasso in dollari è additivo perché ha senso calcolare la somma degli incassi per un certo intervallo di tempo, insieme di prodotti e insieme di negozi
    - ad esempio, in un mese, per una categoria di prodotti e per i negozi in un'area geografica
  - l'additività è una proprietà importante: le applicazioni del data warehouse devono spesso combinare i fatti descritti da molti record di una tabella fatti
    - il modo più comune di combinare un insieme di fatti è di sommarli (se questo ha senso)
    - è possibile anche l'uso di altre operazioni (ad esempio media pesata)

Semi additività e non additività

- I fatti possono essere anche
  - semi additivi
    - se ha senso aggregarli solo rispetto ad alcune dimensioni
      - il numero di pezzi in deposito di un prodotto è sommabile rispetto alle categorie di prodotto e ai magazzini, ma non rispetto al tempo
  - non additivi
    - se non ha senso aggregarli

Interrogazioni di schemi dimensionali

- Gli attributi delle tabelle dimensione sono il principale strumento per l'interrogazione del data warehouse
  - gli attributi delle dimensioni vengono usati per
    - selezionare un sottoinsieme dei dati di interesse
      - vincolando il valore di uno o più attributi
      - ad esempio, le vendite nel corso dell'anno 2000
    - raggruppare i dati di interesse
      - usando gli attributi come intestazioni della tabella risultato
      - ad esempio, per mostrare le vendite per ciascuna categoria di prodotto in ciascun mese

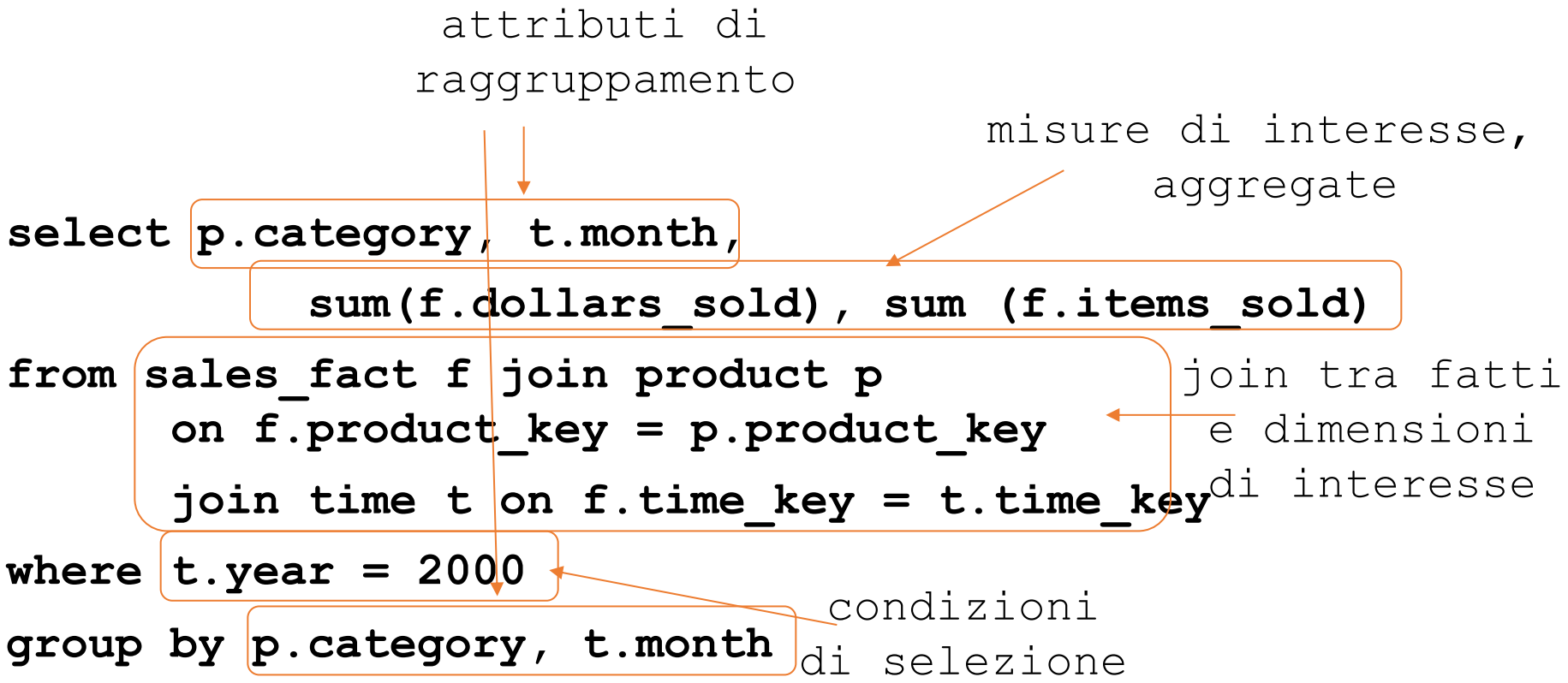
Attributi e interrogazioni

- Dati restituiti dall'interrogazione
  - somma degli incassi in dollari e delle quantità vendute
  - per ciascuna categoria di prodotto in ciascun mese
  - nel corso dell'anno 2000

(product) category	(time) month	(sum of) dollars_sold	(sum of) units_sold
Drinks	gennaio 2000	21.509,05	23.293
Drinks	febbraio 2000	19.486,93	22.216
Drinks	marzo 2000	21.986,43	23.532
Food	gennaio 2000	86.937,77	55.135
Supplies	gennaio 2000	21.554,17	13.541

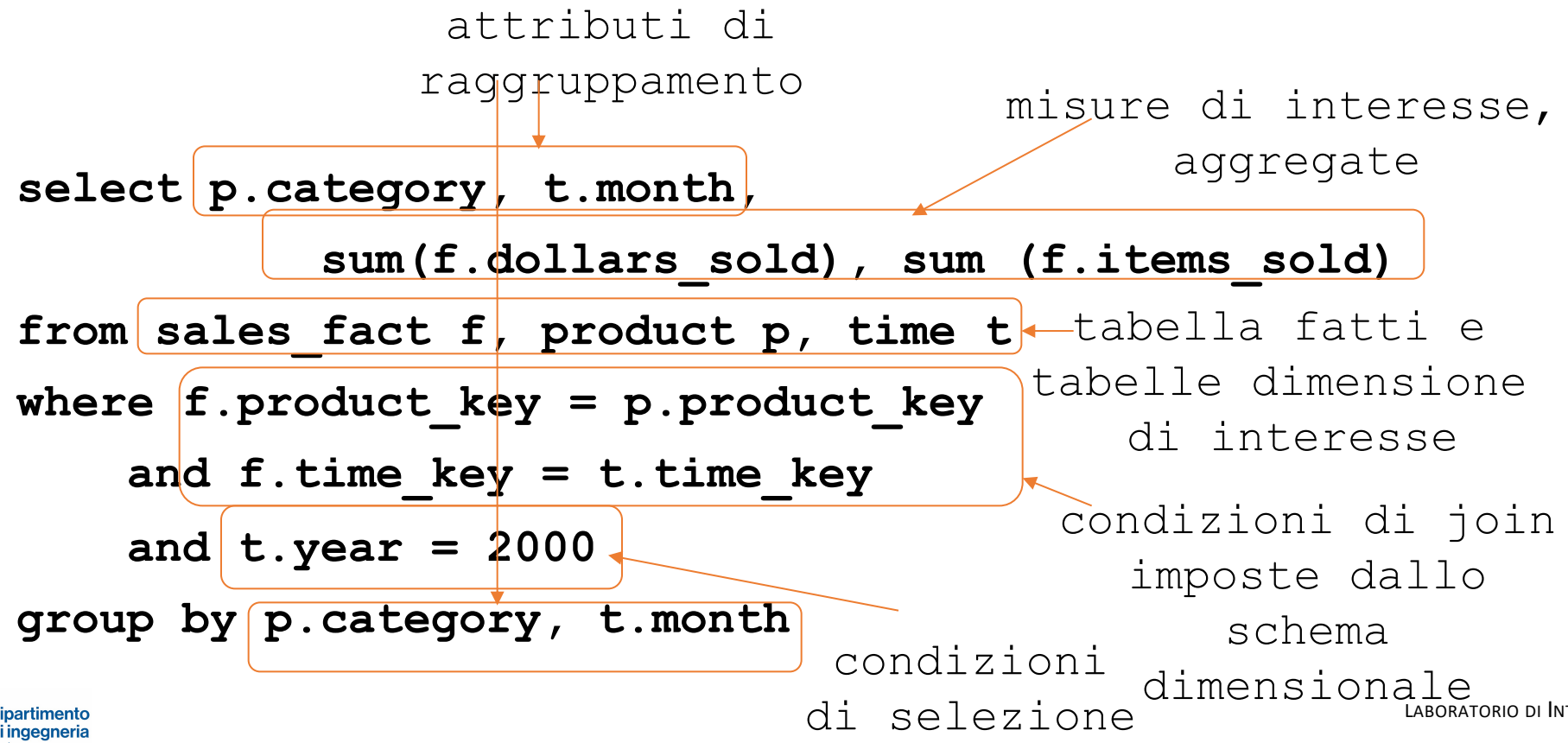
Formato delle interrogazioni

- Le interrogazione assumono un formato abbastanza standard



Formato delle interrogazioni (2)

- Idem, senza join esplicito



Drill down

- L'operazione di drill down aggiunge dettaglio ai dati restituiti da una interrogazione
  - il drill down avviene aggiungendo un nuovo attributo nell'intestazione di una interrogazione e nel raggruppamento
  - diminuisce la grana dell'aggregazione

(product) category	(time) month	(sum of) dollars_sold	(sum of) units_sold

drill down

(product) category	(time) month	(store) city	(sum of) dollars_sold	(sum of) units_sold

Drill down

- Aggiungiamo raggruppamenti (e join)

```
select p.category, t.month, s.city,
            sum(f.dollars_sold), sum (f.items_sold)
from sales_fact f join product p
      on f.product_key = p.product_key
      join time t on f.time_key = t.time_key
      join store s on f.store_key = s.store_key
where t.year = 2000
group by p.category, t.month, s.city
```

Roll up

- L'operazione di roll up riduce il dettaglio dei dati restituiti da una interrogazione
  - il roll up avviene rimuovendo un attributo dall'intestazione di una interrogazione e dal raggruppamento
  - aumenta la grana dell'aggregazione

(product) category	(time) month	(sum of) dollars_sold	(sum of) units_sold

roll up

(product) category	(sum of) dollars_sold	(sum of) units_sold

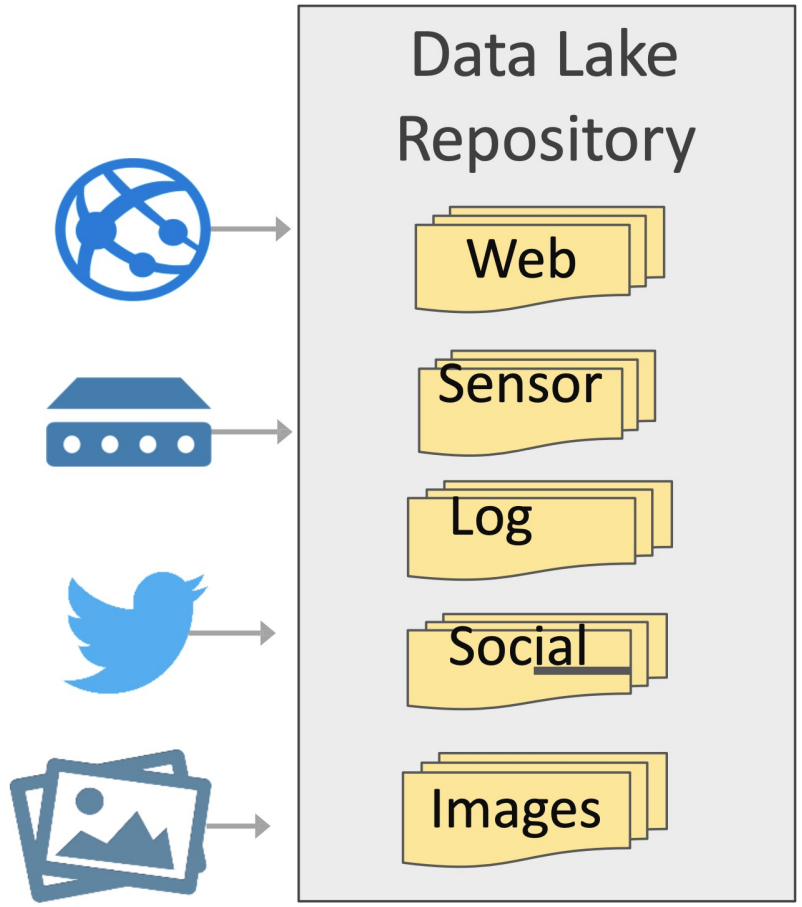
Roll up

- Eliminiamo raggruppamenti

```
select p.category, t.month,
            sum(f.dollars_sold), sum (f.items_sold)
from sales_fact f join product p
      on f.product_key = p.product_key
      join time t on f.time_key = t.time_key
where t.year = 2000
group by p.category, t.month
```

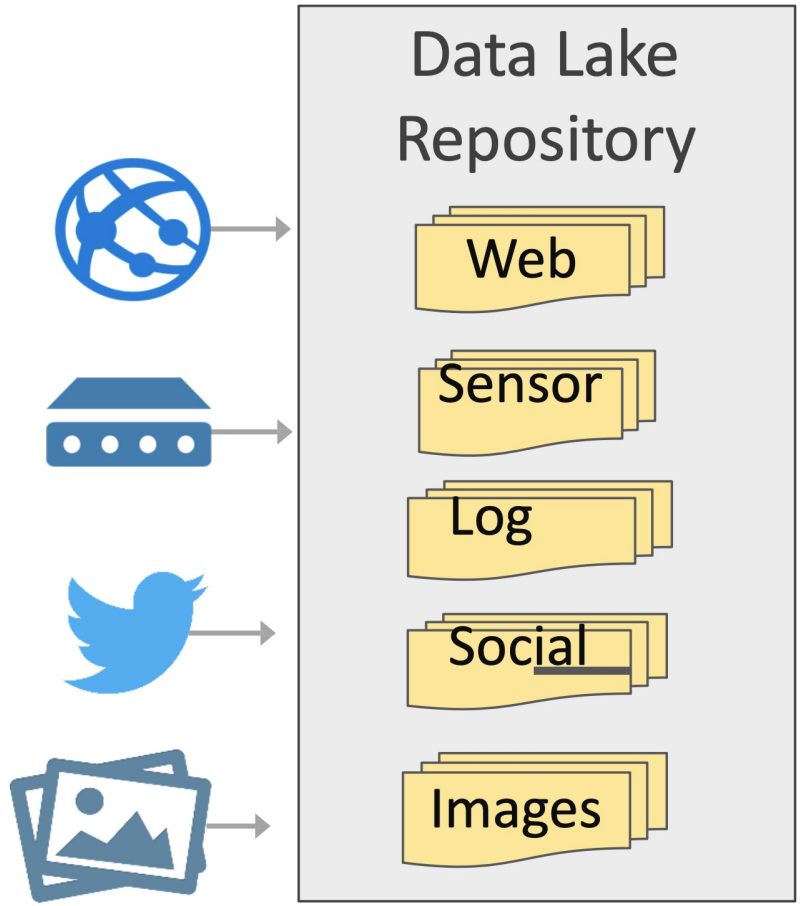
Data Lake

- Un Data Lake è un repository per dati eterogenei nel loro formato nativo
  - Nessuna necessità di imporre *schemi* ai dati (con il significato di schema logico dei classici DB)
  - Facilità di acquisizione dei nuovi dati
  - Gestione di grandi volumi di dati multi-strutturati



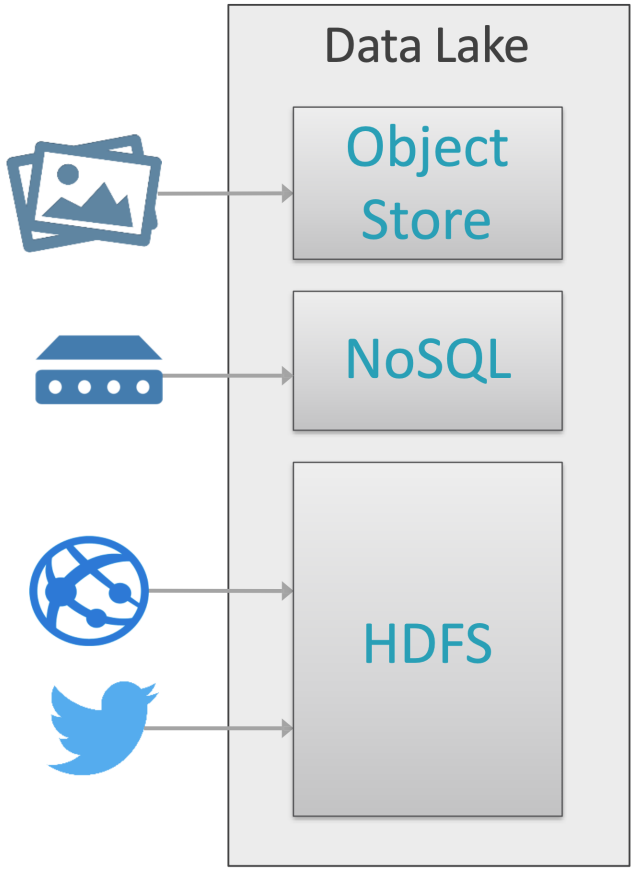
Data Lake

- Un Data Lake è un repository per dati eterogenei nel loro formato nativo
  - Supporto semplice alle funzionalità Data Analytics
  - Velocizzazione dei processi di supporto alle decisioni
  - Gestione semplice per nuovi tipi di dati



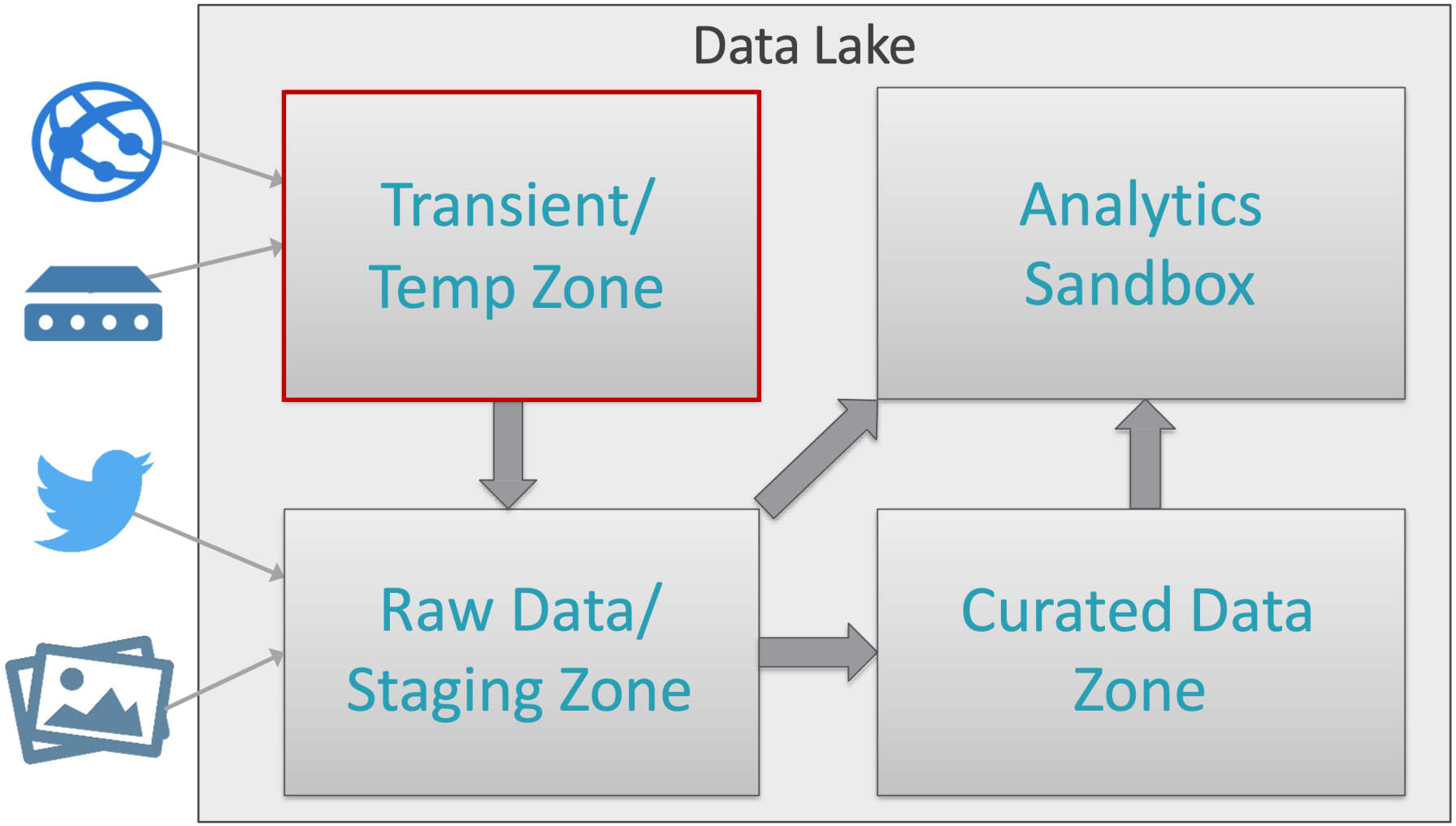
Data Lake

- Implementazione tramite tecnologie Big Data
  - Object Store per dati multimediali (Amazon S3 o Azure Blob Storage)
  - Database NoSQL per dati provenienti da sensori
  - *Hadoop Distributed File System* (HDFS) anche su cluster per i dati social e web



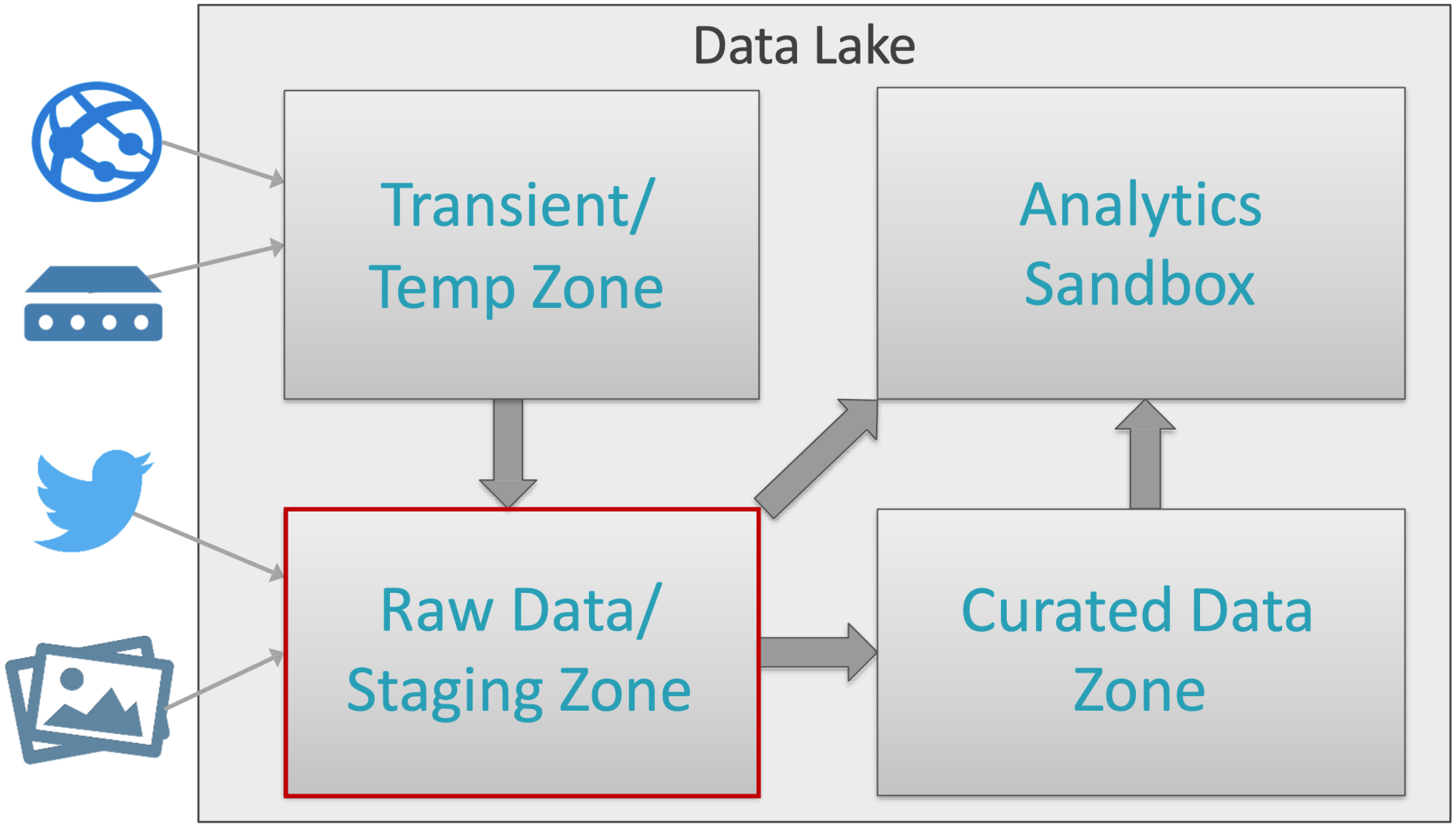
Zone in un Data Lake

- Transient/ Temp Zone
  - Si usa se sono necessari dei controlli di consistenza e/o validità del dato *prima* di entrare nella Raw Data Zone



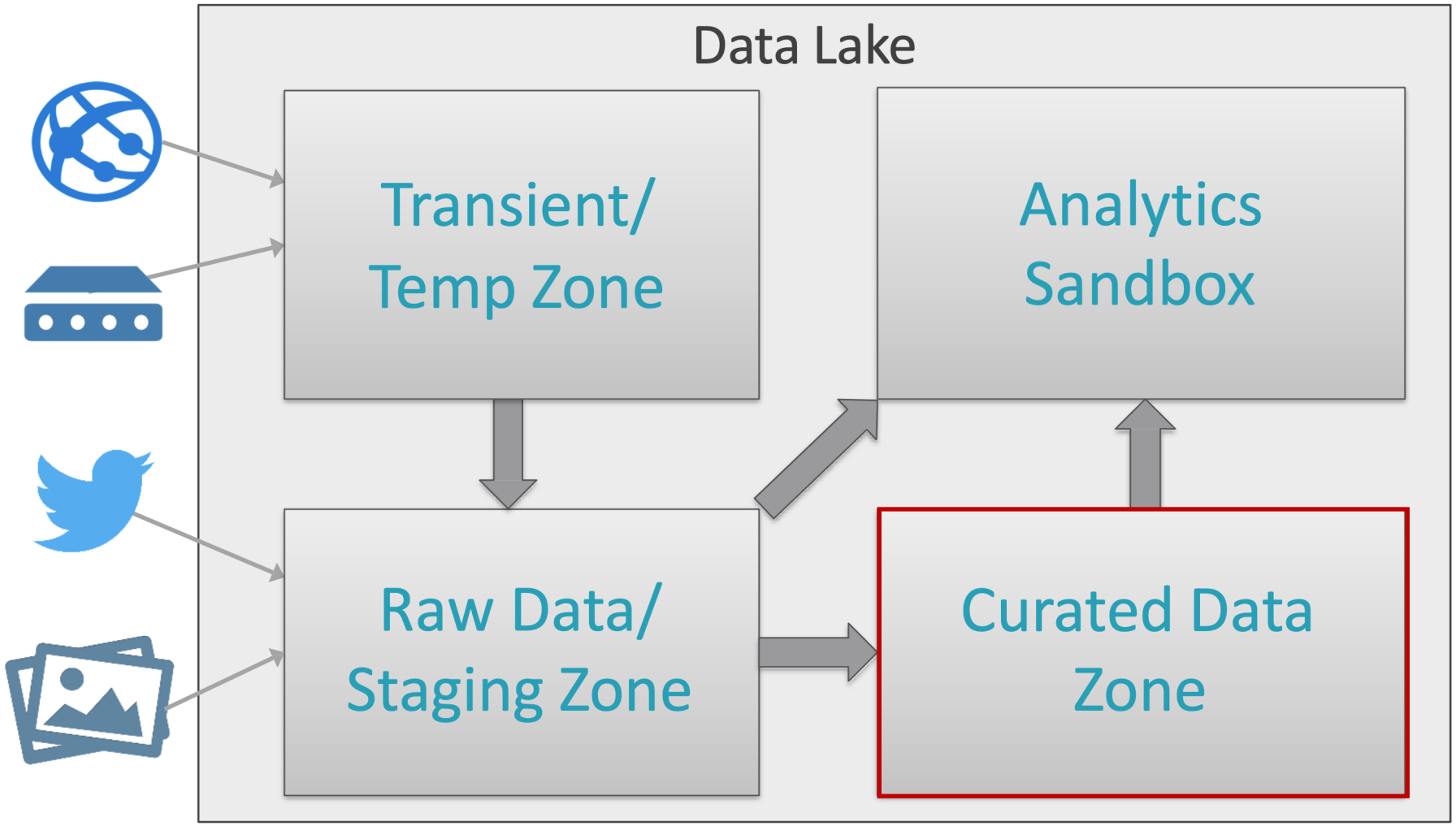
Zone in un Data Lake

- Raw Data/ Staging Zone
  - Supporta qualunque tipo di dato
  - Immutabile
  - Conserva la storia del dato



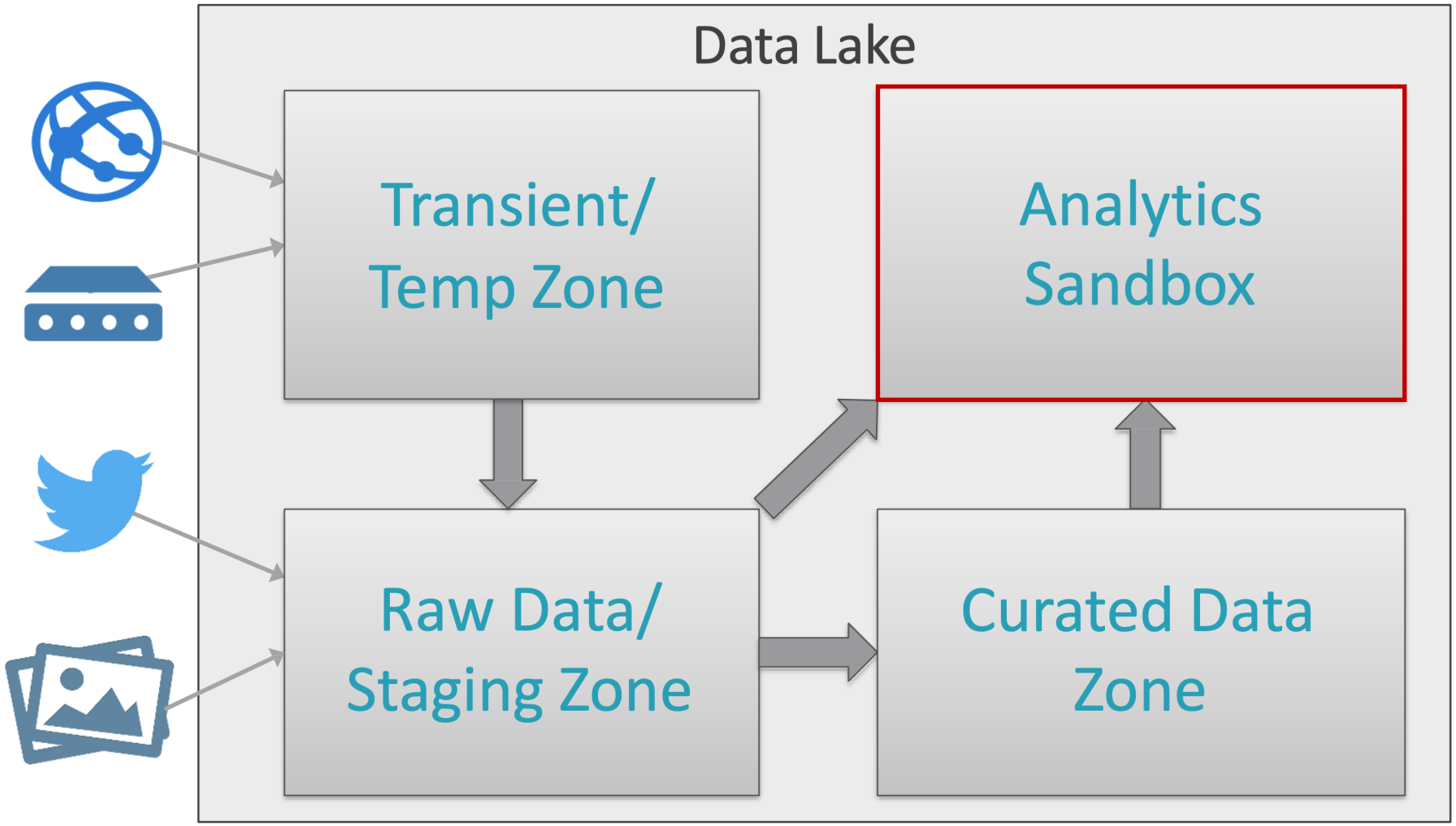
Zone in un Data Lake

- Curated Data Zone
  - Dati organizzati per consentire l'accesso ad altri sistemi o ad altri componenti dell'architettura
  - Implementa criteri di sicurezza standard

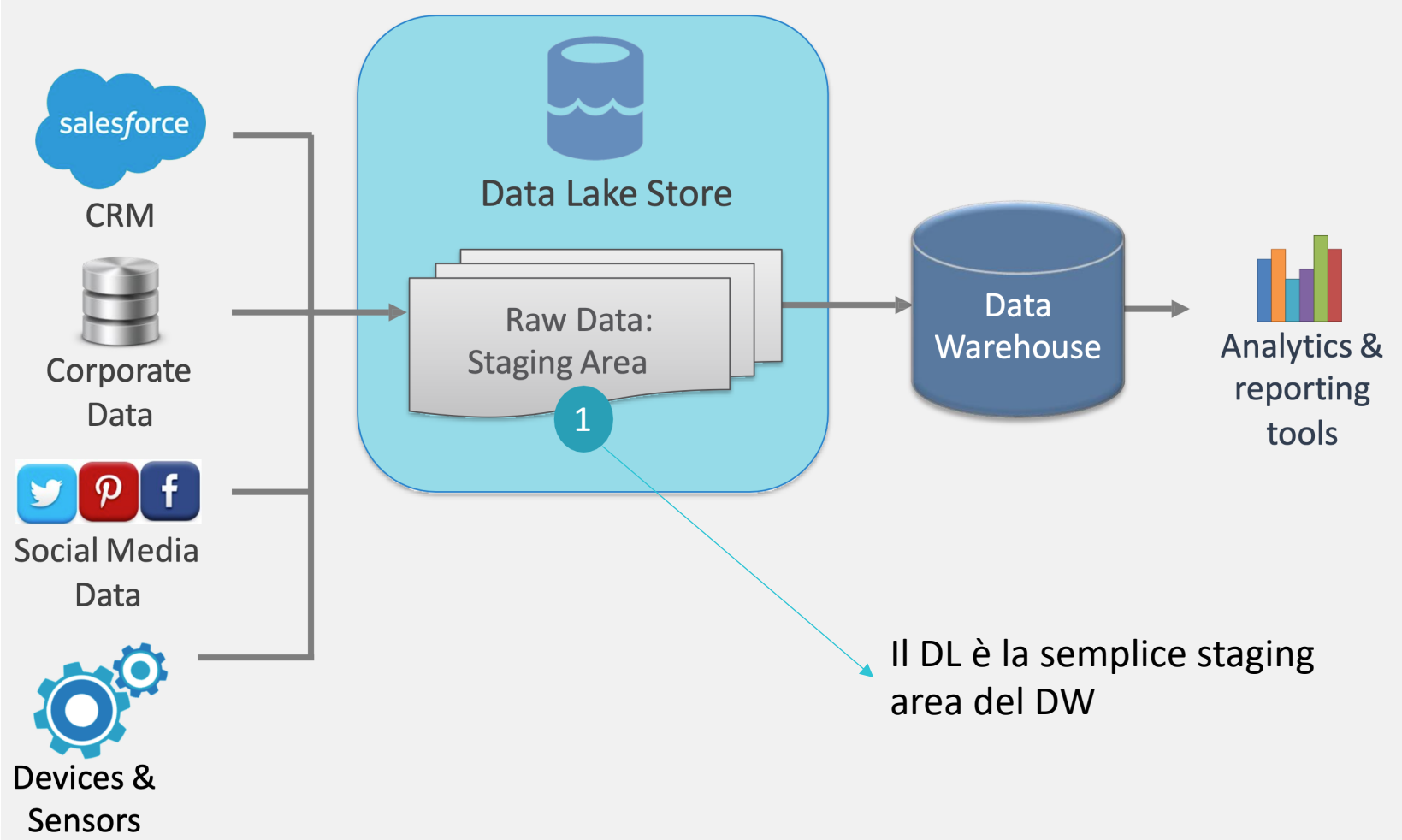


Zone in un Data Lake

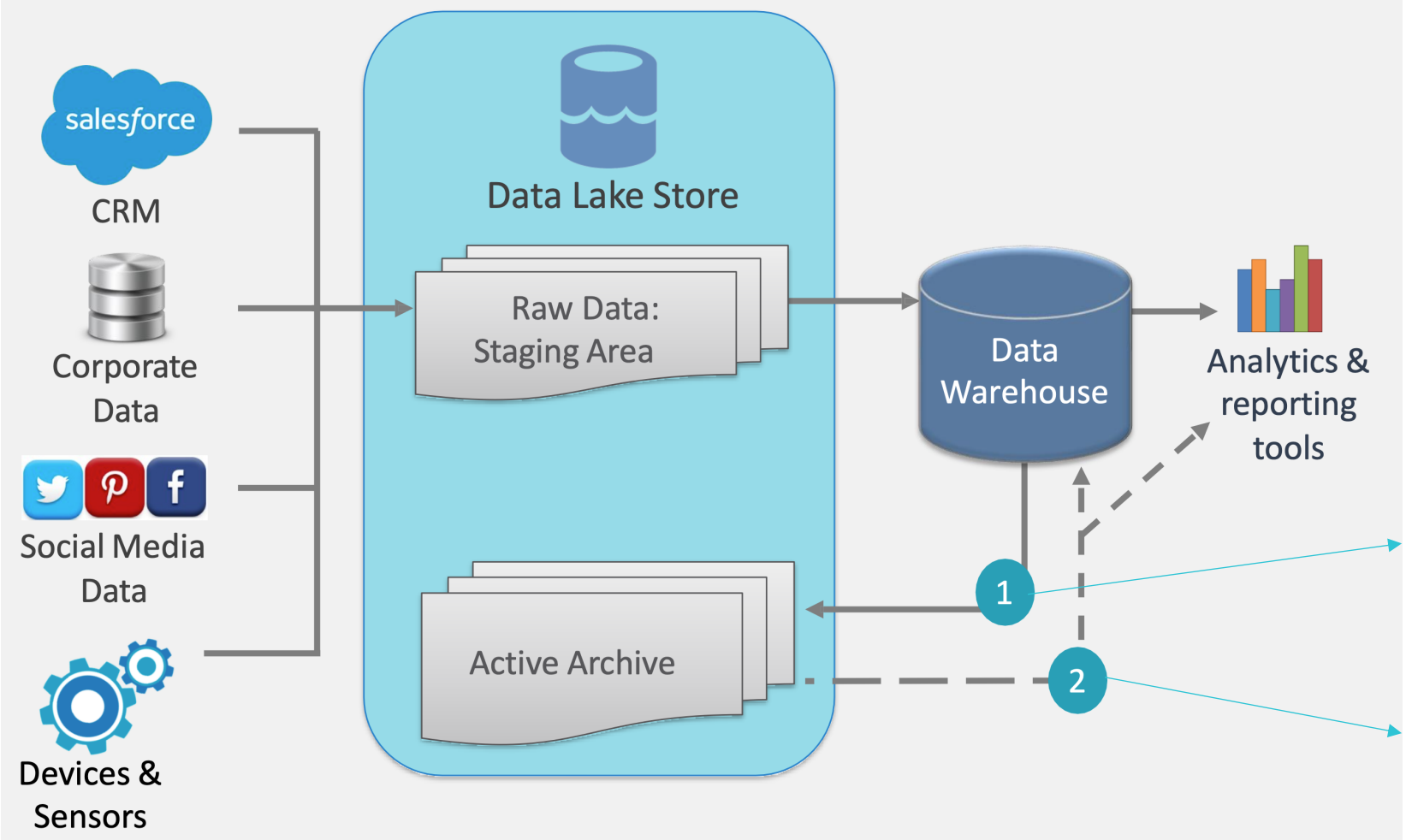
- Analytics Sandbox
  - Zona per l'esecuzione di processi di esplorazione o di data analysis
  - Gestione minimale



Integrazione Data Lake e Data Warehouse



Integrazione Data Lake e Data Warehouse



Integrazione Data Lake e Data Warehouse

Architettura Lambda:

Il DW agisce da serving layer integrandosi con al Curated Data Zone del DL