

La misurazione dei sistemi di Data Warehouse

Luca Santillo (CFPS)

AIPA, 17/5/01

Sommario

- ✦ Introduzione
- ✦ Definizioni & Modelli
- ✦ FP & Data Warehouse?
- ✦ Stima dell'impegno
- ✦ Conclusioni

Introduzione

✦ Ipotesi

- ✦ I metodi di Misurazione Funzionale del Software (IFPUG, COSMIC, NESMA, ...) sono indipendenti dalle tecniche e dalle tecnologie di implementazione

✦ Considerazioni

- ✦ Contesti speciali (es. Data Warehouse) possono fare uso di specifiche tecnologie.

E

- ✦ Possono assumere specifiche tipologie di utente (“punto di vista”) & specifici approcci all’analisi (“modelli”).

Introduzione

✦ In pratica

- ✦ I concetti generali, le definizioni e le regole di conteggio possono essere difficili da applicare a contesti specifici (es. Data Warehouse).

✦ L'approccio proposto

- ✦ Istanza di un metodo generale per la misurazione e la stima di impegno (non un metodo “alternativo”).
- ✦ Linee Guida per la misura e la stima dei sistemi di Data Warehouse.

Definizioni & Modelli

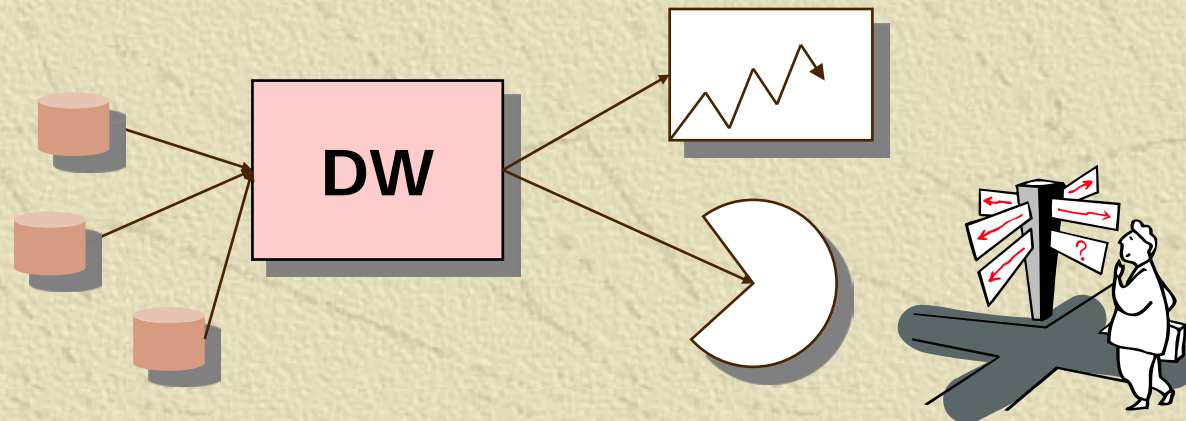
- ✦ Data Warehousing
- ✦ Enterprise DW & Data Mart
- ✦ Fasi del Flusso dei Dati (ETL, OLAP)
- ✦ Cubo Multi-dimensionale (modello dati)
- ✦ Star Schema (modello dati)
- ✦ Meta-dati
- ✦ IFPUG Function Point
- ✦ COSMIC Full FP

Data Warehousing

✦ Raccolta di un insieme di dati che sia:

- ✦ Subject-oriented
- ✦ Integrato
- ✦ Time-variant
- ✦ Non volatile

...in supporto
all'analisi del business &
al processo decisionale.



Confronto tra tipologie

✦ MIS vs DW

	Sistemi Operazionali	Sistemi Data Warehouse
Scopo	Operazioni routinarie, Fotografia del presente	Reperimento & Analisi dati, Mantenimento dello storico
Struttura	DB ottimizzato per data entry e query singola	DB ottimizzato per data query globale
Modello dati	Normalizzato	Multi-dimensionale
Natura dei dati	Dettaglio	Sintesi & Dettaglio
Dati duplicati	No (DB normalizzato)	Sì (DB de-normalizzato)
Dati Derivati	Raramente	Comuni
Tipo Utente	Basso livello / Operativo	Alto livello / Manageriale

Peculiarità del DW

✦ Il DW possiede specifici:

- ✦ Tipi di utente
 - *Top level, Decision-makers*
- ✦ Modelli di analisi (dati)
 - *Star Schema, Cubo Multi-dimensionale*
- ✦ Tecnologie
 - *Elevata dimensione fisica dei dati*
 - *Performance*
- ✦ Tools per l'implementazione
 - *Personalizzazione di Pacchetto più che Sviluppo*

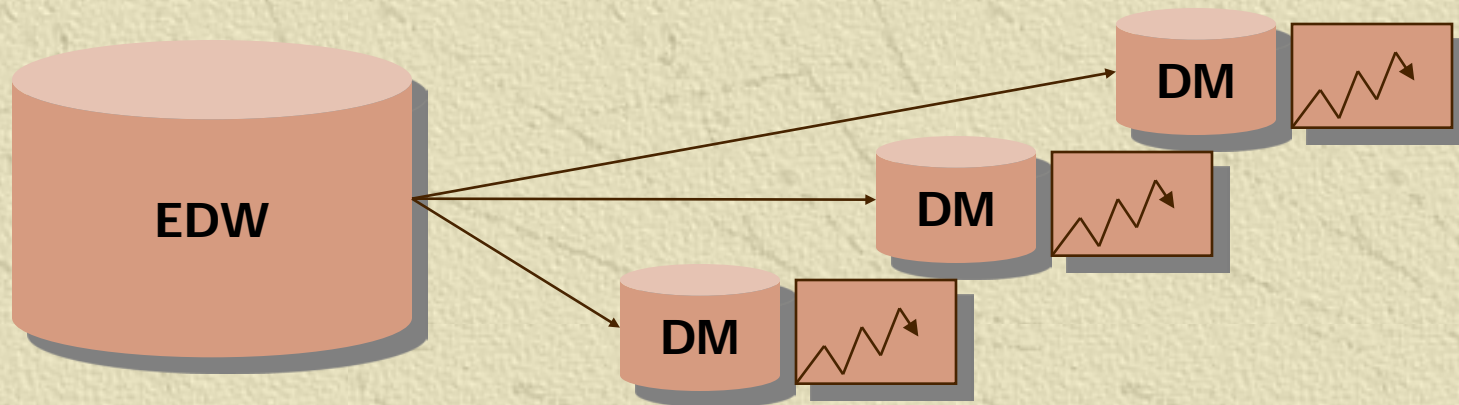
Enterprise DW & Data Marts

✦ EDW

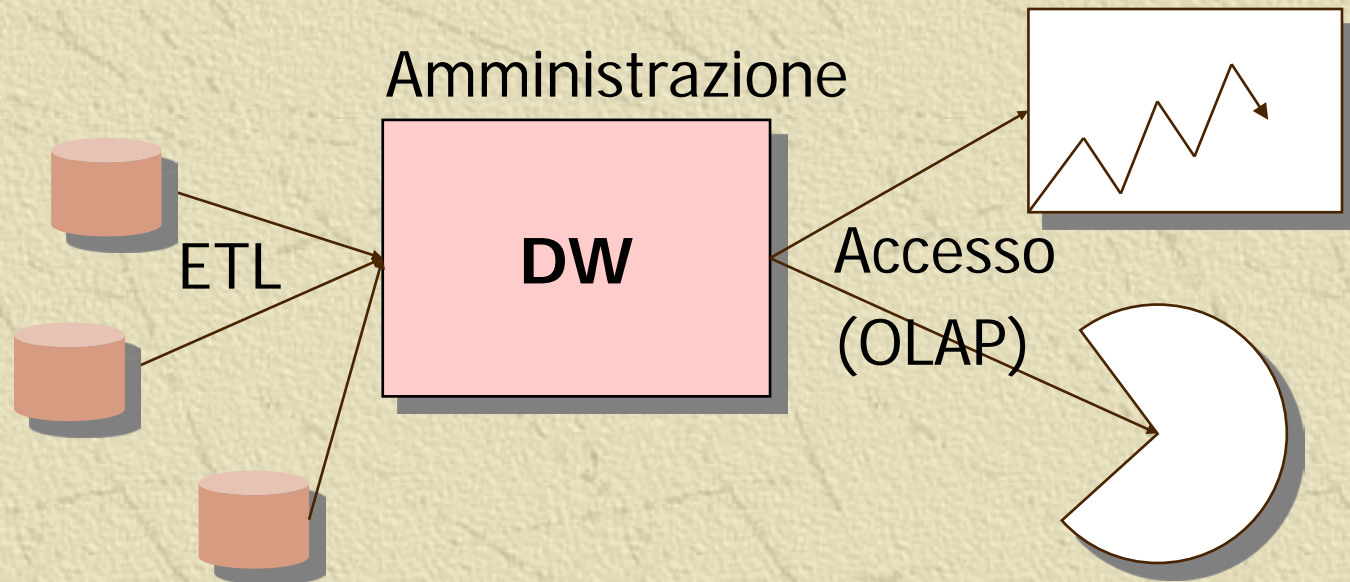
- ✦ Raccolta di dati per l'intera organizzazione

✦ DM

- ✦ DW tematico (“subject-specific”)
 - Può essere *dipendente* o *indipendente*



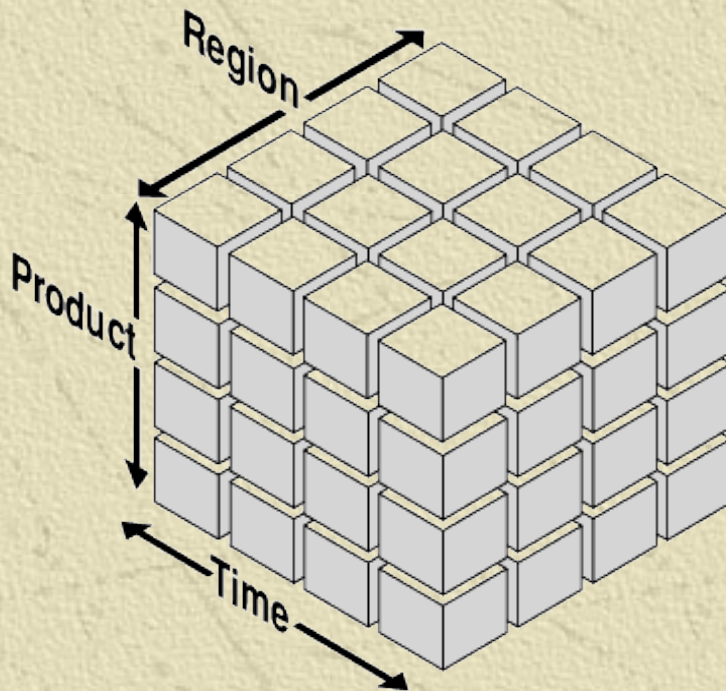
Fasi del Flusso dei Dati



ETL = Extraction, Transformation, Loading

OLAP = On-Line Analytical Processing

Cubo Multi-dimensionale



✦ Fatti

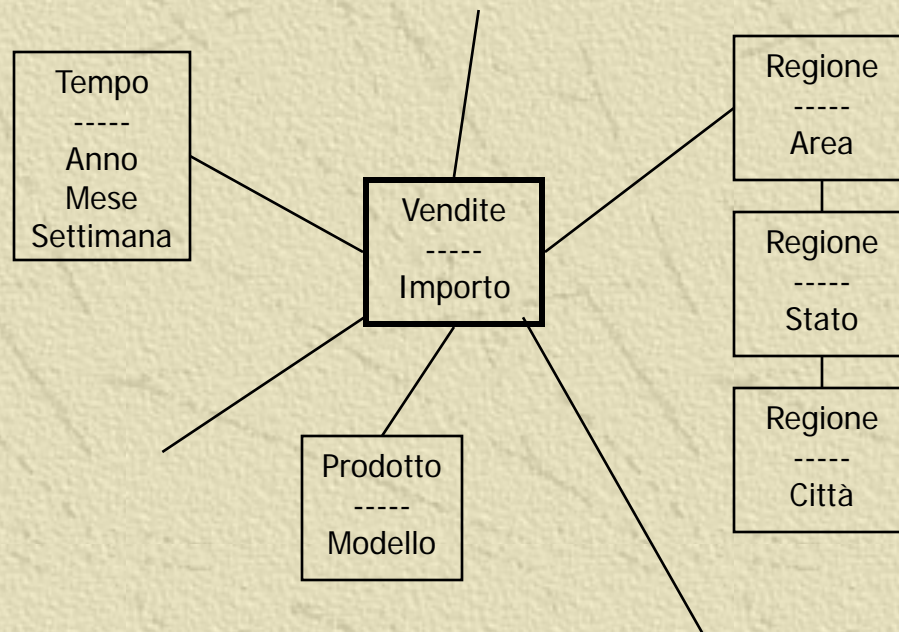
- ✦ Misure, o valori (Quanto)

✦ Dimensioni

- ✦ Categorie, o attributi (Chi, Cosa, Dove, Quando)
- ✦ Permettono lo “slice & dice” del cubo

Star Schema

- ✦ Istanza di un cubo
- ✦ È composto da dimensioni gerarchizzate
- ✦ Permette il “drill up & down” sui Fatti

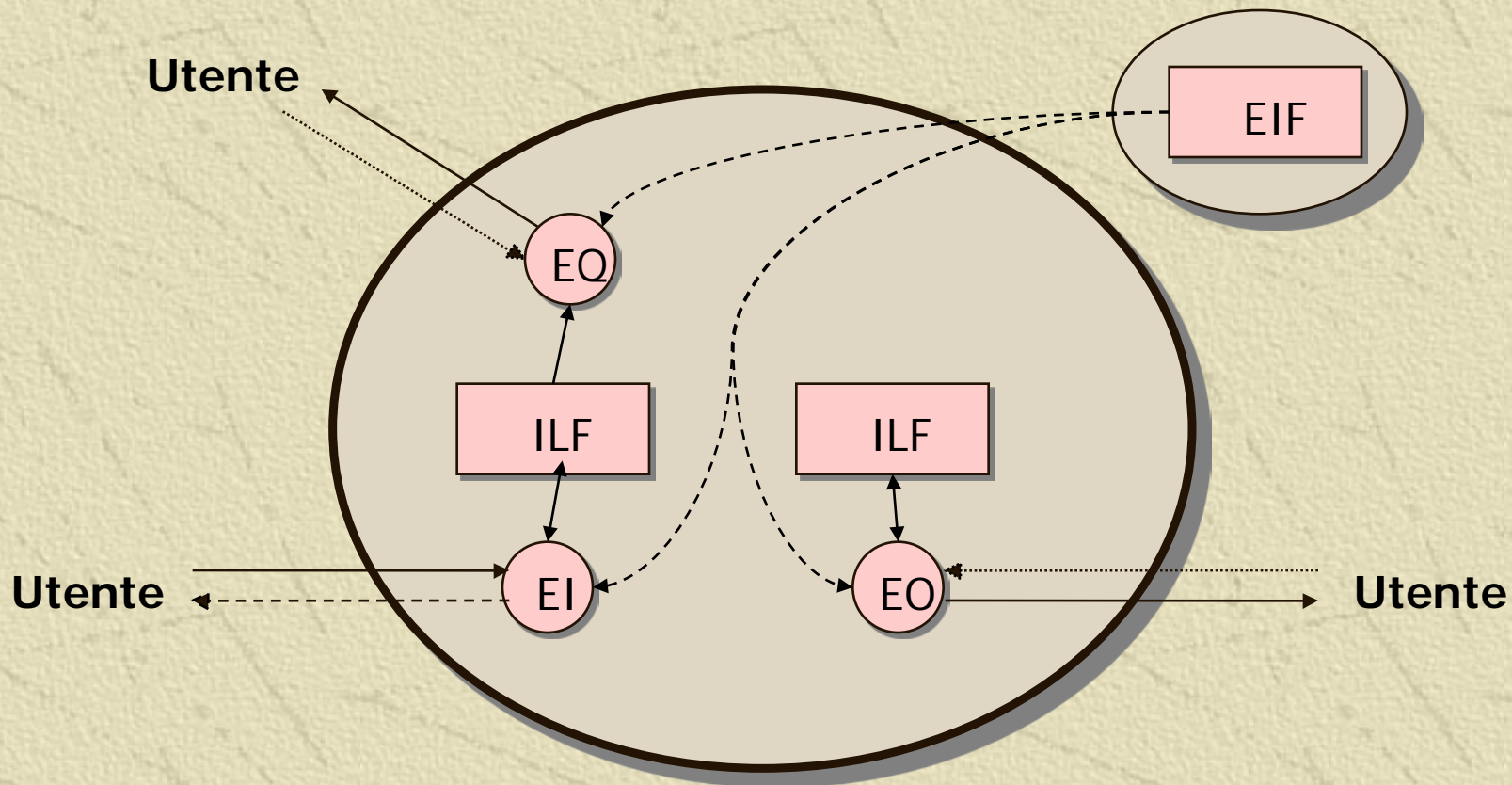


Meta-dati

✦ Dati riguardanti i dati

- ✦ I meta-dati tengono traccia di “cosa” è “dove” nel DW:
 - *Profili utente*
 - *Statistiche d'uso*
 - *Mapping fisico-logico dei dati*
 - ...

IFPUG Function Point (1)

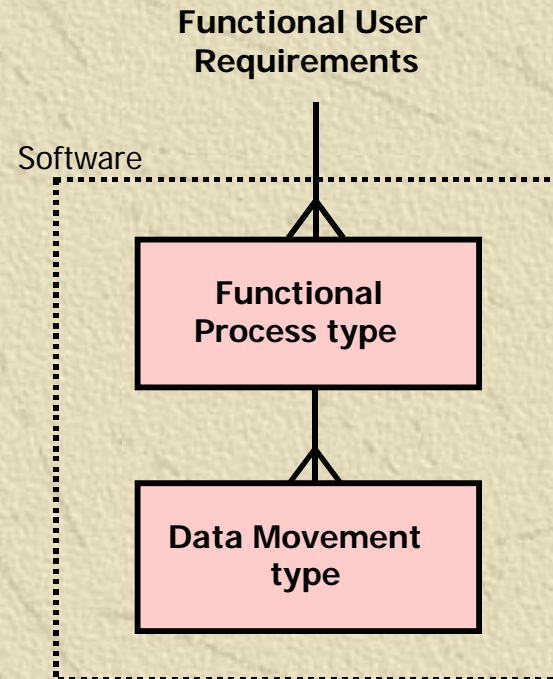


IFPUG Function Point (2)

- ✦ La complessità dei file dipende da:
 - ◆ Attributi dati (contenuti) (DET)
 - ◆ Sottogruppi di attributi dati (RET)
- ✦ La complessità delle transazioni dipende da:
 - ◆ Attributi dati (elaborati) (DET)
 - ◆ File referenziati (letti/scritti) (FTR)
- ✦ La complessità si converte in Unadjusted FP
- ✦ Si applica il Value Adjustment Factor (VAF)
 - ◆ Per includere requisiti tecnici e sulla qualità

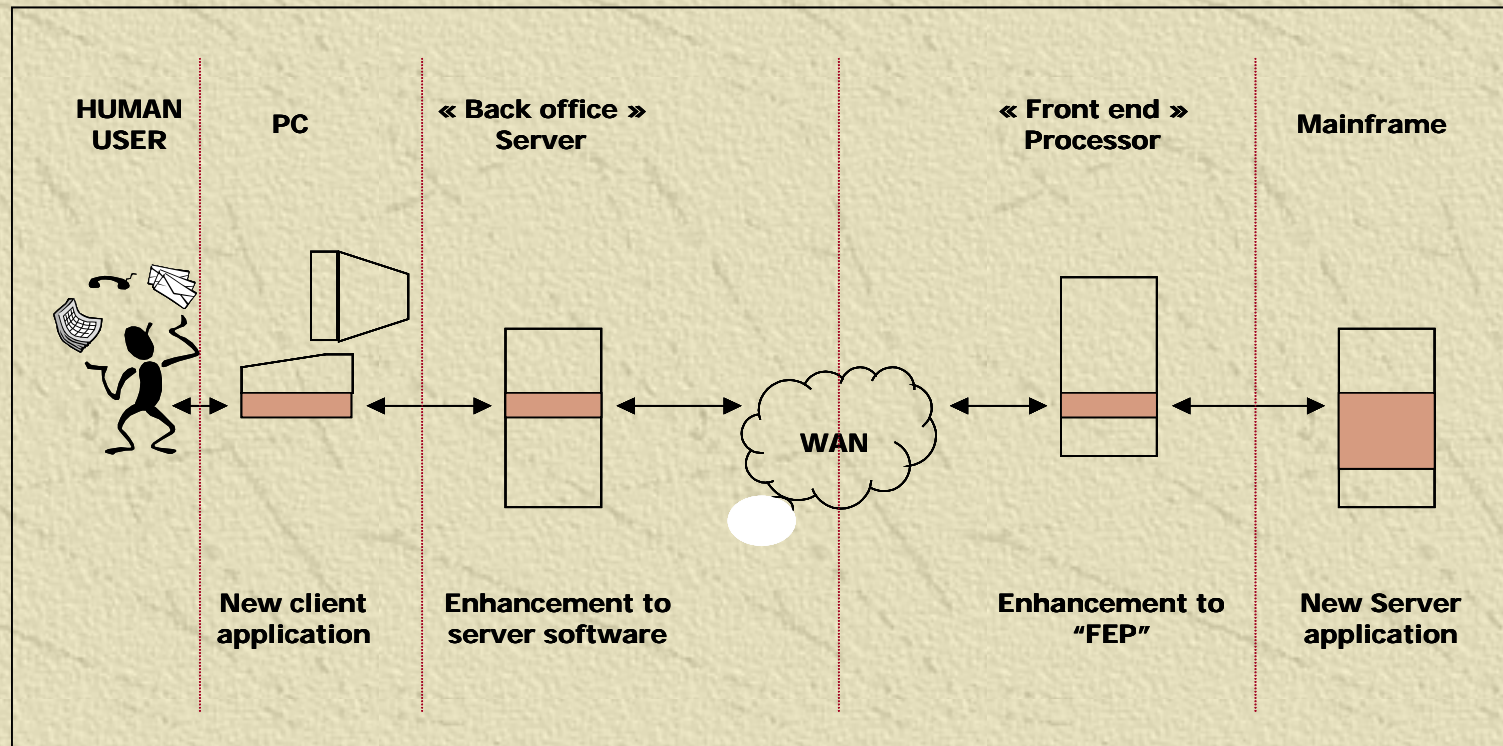
COSMIC Full FP (1)

- ✦ Componente dati non conteggiata
- ✦ Nessun limite superiore alla complessità (size) del singolo processo
 - ✦ Somma flussi (Entries, Exits, Reads, Writes)
- ✦ Nessun VAF



COSMIC Full FP (2)

✦ Si distinguono gli strati (“layers”) secondo la vista utente e/o le necessità gestionali



FP & Data Warehouse?

- ✦ Relazione EDW / DMs
- ✦ Peculiare rappresentazione logica dei dati (e loro aggregazione dal punto di vista utente)
- ✦ Gerarchie dimensionali dei dati
- ✦ Meta-dati
- ✦ Procedure ETL
- ✦ Procedure di accesso (OLAP)
- ✦ ...

FP & Data Warehouse!

✦ È possibile un “mapping” tra concetti della misurazione funzionale e caratteristiche dei sistemi data warehouse, ma...

✦ ...occorre specificare alcuni criteri

- ✦ Linee Guida

✦ Non occorrono metriche alternative

Punto di vista dell'utente & Confine

✦ Punto di vista dell'utente

- ✦ Composizione di vari ruoli/viste
 - Utente finale
 - Amministratore per fase (ETL, DB_{DW}, Accesso)
 - Sistemi sorgente

✦ Confine dell'Applicazione

- ✦ Coerente con la struttura organizzativa (dipartimenti/settori)
- ✦ Autonomia di gestione dell'EDW da ogni DM
- ✦ Autonomia di gestione di un DM da ogni altro DM
- ✦ Ri-definizione dei confini in casi eccezionali

Funzioni Dati: Sorgenti Esterne

✦ EIFs per l'EDW (o DM_{INDIP})

- ✦ Distinti come ILF dei sistemi esterni
 - ✦ Strutturati in DET/RET come da vista utente del DW
- ✦ I duplicati fisici su aree/piattaforme differenti non sono conteggiati.
- ✦ Nessun EIF, se flussi di dati “comandati” dal sistema sorgente (EO sorgenti)

Funzioni Dati: Dati Interni

- ✦ 1 ILF per ogni “stella logica” distinta
- ✦ Struttura RET secondo le dimensioni e le misure
 - ◆ Fact Table
 - ◆ N Tabelle dimensionali al 1° ordine (es. “area”, “nazione”, “città” = 1 RET “regione”).
- ✦ Linee guida specifiche per i DET
 - ◆ Non tutti i dati pre-calcolati sono riconosciuti dall’utente.
 - ◆ DET ricorsivi su tabelle gerarchiche.

Funzioni Dati: Meta-dati

✦ *Meta-dati tecnici* solitamente non contati:

- ✦ Es: Mappatura fisico-logica dei dati

✦ Alcuni riconoscibili dall' "utente":

- ✦ Es: Archivio dei profili utente
- ✦ Es: Statistiche d' utilizzo del sistema

✦ *Meta-dati di business* solitamente contati:

- ✦ Es. "*Dizionario dei dati*"
- ✦ Informazioni sulla proprietà dei dati

Transazioni: ETL

- ✦ 1 solo EI per ogni ILF di tipo “stella logica”
 - ✦ *Lettura* file esterni sorgenti + *pulizia & trasformazione* contenuti + *eventuale lettura* meta-dati + *caricamento* in tabelle destinarie: un unico processo dal punto di vista dell’utente.
- ✦ FTR: ILF di destinazione + EIFs sorgenti necessari + eventuali meta-dati.
- ✦ DET: in accordo con i DET degli FTR.

Transazioni: Amministrazione

- ✦ La fase di amministrazione è composta da processi di tipo tradizionale sui meta-dati.
 - ✦ CRUD (Create, Read, Update, Delete).
 - ✦ Misurata come i sistemi “tradizionali”.

Transazioni: Accesso (OLAP)

✦ (Almeno) 1 processo (EO/EQ) per ogni ILF di tipo “stella logica”

- ✦ “Drill down/up” su 1 stella non costituisce differente output (equivalente a filtro di query).
- ✦ Il meccanismo in sé per il “drill” è solitamente fornito dai tool come una listbox su ogni attributo “drillabile” = 1 EQ di bassa complessità (per ogni distinto attributo di ogni distinta stella).

VAF e Calcolo FP Finale

- ✦ VAF non considerato (o sempre = 1)
 - ✦ ISO 14143 (Functional Measurement Methods).
 - ✦ La teoria della misurazione dimostra l'inadeguatezza del VAF.
 - ✦ Requisiti tecnici e di qualità inclusi nella fase successiva di stima dell'impegno.
- ✦ Il risultato finale della misura è la somma degli Unadjusted FP.

Stima dell'Impegno

- ✦ Numerosi aspetti peculiari di produttività per DW:
 - ✦ “Cut-off” della complessità delle funzioni
 - Matrici IFPUG: complessità Bassa, Media, Alta.
 - ✦ Il riuso (interno ed esterno) è rilevante.
 - ✦ Attenzione alla strutturazione dei dati del DW ha impatto rilevante (negativo) sulla fase di analisi.
 - ✦ Tools e tecnologie specifiche hanno impatto rilevante sulla fase di *implementazione*,
 - In particolare, per la fase di Accesso (personalizzazione di pacchetti).
 - ✦ Non esistono adeguati benchmarks.

Modello di Stima Strutturato

1. La misura è “aggiustata” per:

- ◆ Classi di intervento/funzioni
 - Specifici coefficienti moltiplicatori per EI_{ETL} , ILF_{Stella} , ...
- ◆ Riuso
 - In termini di comunanza di DET/RET/FTR tra funzioni “simili”.
- ◆ Tecnologia
 - Specifici coefficienti moltiplicatori per $EO/EQ_{Accesso}$, EQ_{Drill} , ...

Modello di Stima Strutturato

2. Si applica una formula generica (benchmarking/regressione statistica):

$$\text{Eff}_{\text{avg}} = A \text{ "Size" }^B$$

(o analoga formula)

Modello di Stima Strutturato

3. Si applica un modello à *la COCOMO II*, per calibrare la stima ai fattori “locali” di progetto:

$$\text{Eff} = \text{Eff}_{\text{avg}} \times \prod \text{CD}_i$$

dove i Cost Drivers contemplano solo gli aspetti non implicitamente considerati ai passi precedenti.

Miglioramenti

✦ COSMIC Full FP:

- ✦ Nessun “cut-off” sulla complessità delle funzioni.
- ✦ Il concetto di “layer” permette di applicare valori di produttività differenziati ai vari segmenti.
- ✦ Ma: misurano solo indirettamente la componente dati.
- ✦ Quindi: si attendono valori di produttività estremamente differenti rispetto ai FP IFPUG.

✦ Statistiche e Benchmarking:

- ✦ Le analisi di benchmarking in corso includeranno necessariamente campioni di DW, a causa dell’elevata diffusione in corso di tali sistemi.

Conclusioni

- ✦ È possibile realizzare un set completo di linee guida di conteggio sulla base di elementi tipici dell'analisi dei sistemi DW.
- ✦ Le linee guida sono scalabili (EDW/DM).
- ✦ Le linee guida possono essere formulate e organizzate conformemente alle regole standard.
- ✦ La trasformazione della metrica in Impegno non può prescindere da specifiche differenziazioni per tipologia e tecnologia.
- ✦ Il metodo dei COSMIC FFP si preannuncia più “solido”, eliminando la necessità di alcuni aggiustamenti sulla misura, per la stima d'impegno.

La Fine...?

... è solo l'inizio!

Futuri sviluppi:

- ◆ Evoluzione teorica del metodo,
- ◆ Sperimentazioni sul campo,
- ◆ Benchmarking mirato,
- ◆ Misure COSMIC FFP.



Grazie!

luca.santillo@gmail.com