



Disegno ETL

Un approccio metodologico

Ing. Claudio Sanft

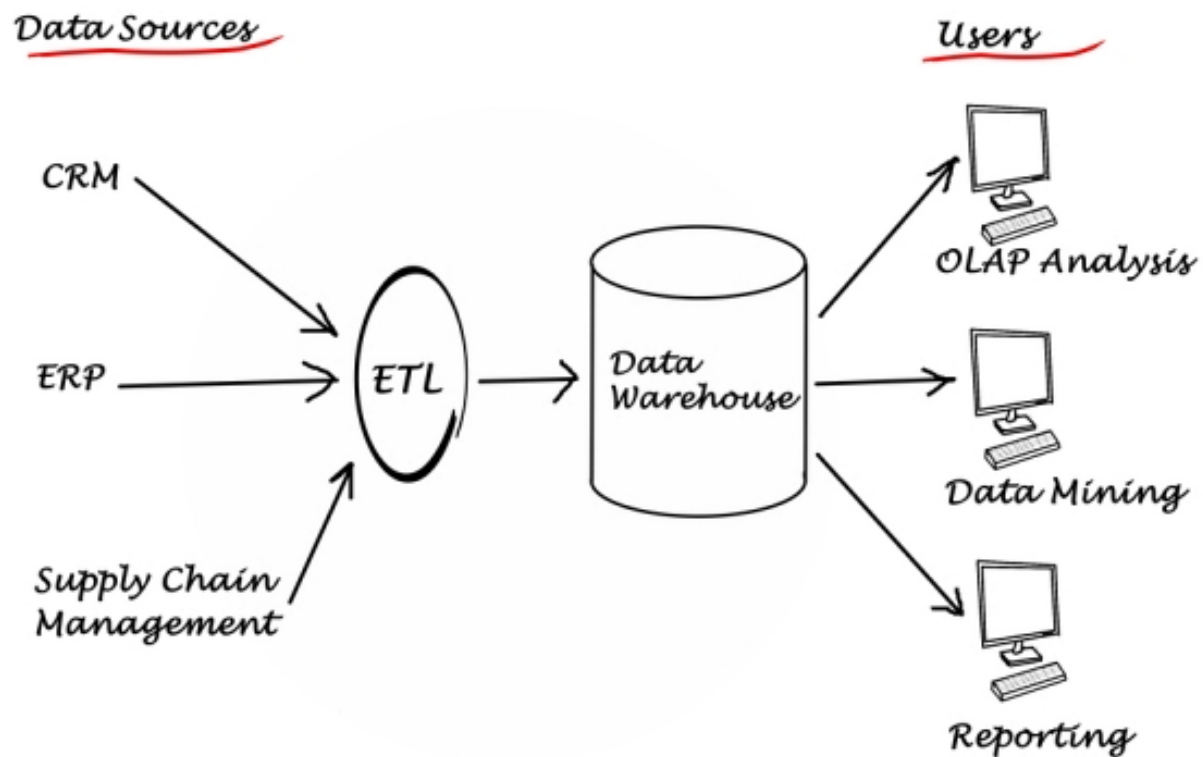
Ordine degli Ingegneri della Provincia di Roma

Roma, 4 Aprile 2025



Il tema

Alimentazione di un Data Warehouse/Data Mart





Obiettivo

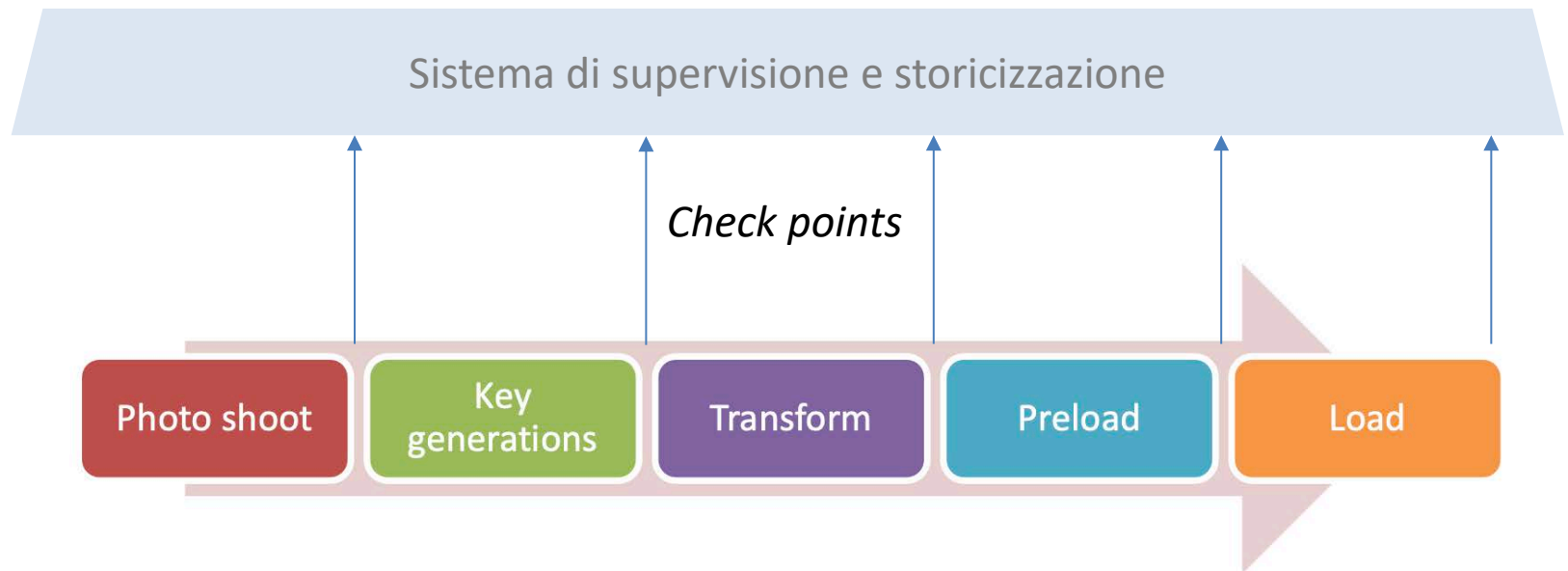
Individuare gli elementi di ripetibilità

Valorizzare la creatività



Schema di sviluppo

Precondizione: identificare la **coordinata di flusso**

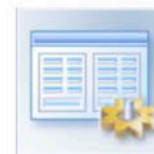




Schema di sviluppo

Sistema di supervisione e storicizzazione

Data inizio: 26/03/2025 09:09:32



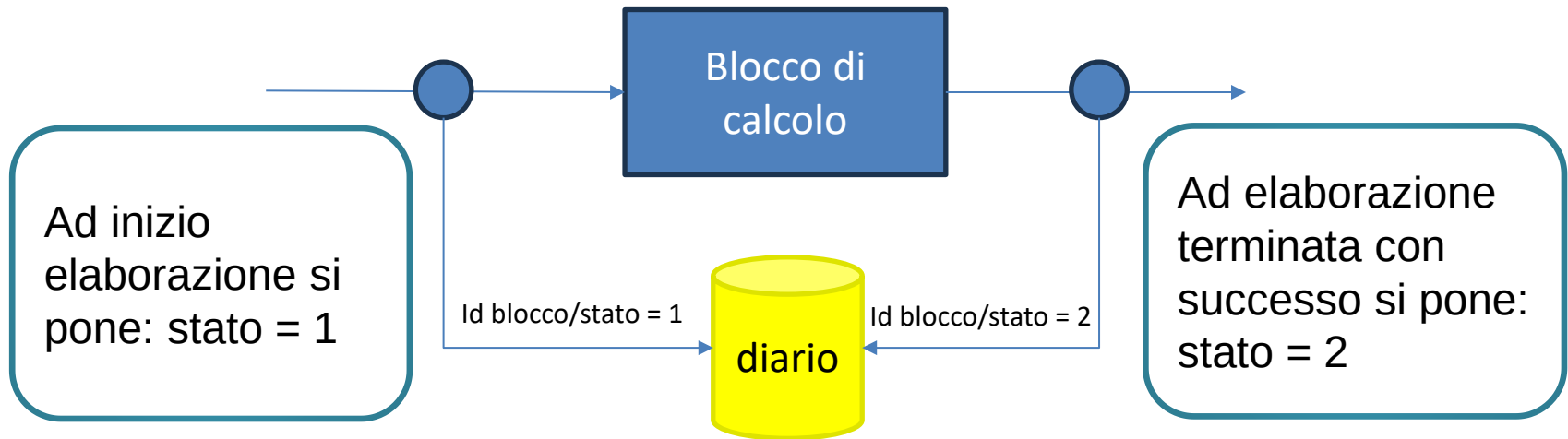
Componente	Run flag componente	Data inizio componente	Data fine componente	Error Number	Error Timestamp	Error Context	Err
STG	✓	26/03/2025 02:27:07	26/03/2025 03:56:50				
STG_1	✓	26/03/2025 02:27:09	26/03/2025 03:23:02				
STG_2	✓	26/03/2025 02:27:09	26/03/2025 03:56:48				
DWH_HIE	✓	26/03/2025 03:56:53	26/03/2025 04:28:37				
DWH_SIO	✓	26/03/2025 03:56:53	26/03/2025 06:56:19				
DWH_AHE_AN_FT_PARALLELO	✓	26/03/2025 03:56:55	26/03/2025 05:01:15				
DWH	⚙	26/03/2025 09:16:38	31/12/2999 00:00:00				
DWH_AHE	⚙	26/03/2025 09:16:40	31/12/2999 00:00:00				
DWH_AHE_FT_MAG	⚙	26/03/2025 09:16:43	31/12/2999 00:00:00				



Dinamica dei check point: impiego

La tabella "diario" conserva gli stati di avanzamento dei blocchi o settori di calcolo

All'avvio, tutti gli stati sono a "0"



Se la procedura termina con successo, tutti gli stati sono impostati a "0"



Dinamica dei check point: utilità

Ad ogni avvio il sistema controlla lo stato dell'ultima esecuzione

Se stato = 1



L'ultima volta vi è stato un ABEND. Si riprende da dove si è interrotto

Se stato = 0



L'ultima volta l'esecuzione ha terminato con successo. Si riparte con un nuovo processamento



Schema di sviluppo

Photo shoot

Preleva i dati dalle sorgenti, astenendosi da qualunque trasformazione. Solo la porzione necessaria di dati viene prelevata, in conformità con la coordinata di flusso



Schema di sviluppo

Procedimento

Key generation

Chiavi naturali → chiavi surrogate

Memorizzazione coppia: Surrogate key/natural key in tabelle di staging area

Vantaggi

Disaccoppiamento chiave sorgente/chave tecnica $\Rightarrow$ robustezza in caso di cambio sorgente;

chiavi naturali complesse $\rightarrow$ chiavi a campo singolo $\Rightarrow$ quadrature molto più semplici;

Arricchimento

Subito dopo aver calcolato le chiavi surrogate, queste vengono aggiunte ai flussi provenienti da Photo shoot, in modo da prepararli per la fase successiva.



Schema di sviluppo

Transform

il blocco più creativo dell'intero sistema. In esso, i dati vengono elaborati per conformarsi allo schema DWH finale.



Schema di sviluppo

Preload

Il flusso dallo stadio "Transform" alimenta le **Mirror Tables**: tabelle a svuotamento, il cui tracciato record replica quello delle tabelle target, addizionato di un campo che indica l'operazione di caricamento per ciascun record

Mirror Table = Target table + operation field (Insert, **U**ppdate, **D**elete)



Schema di sviluppo

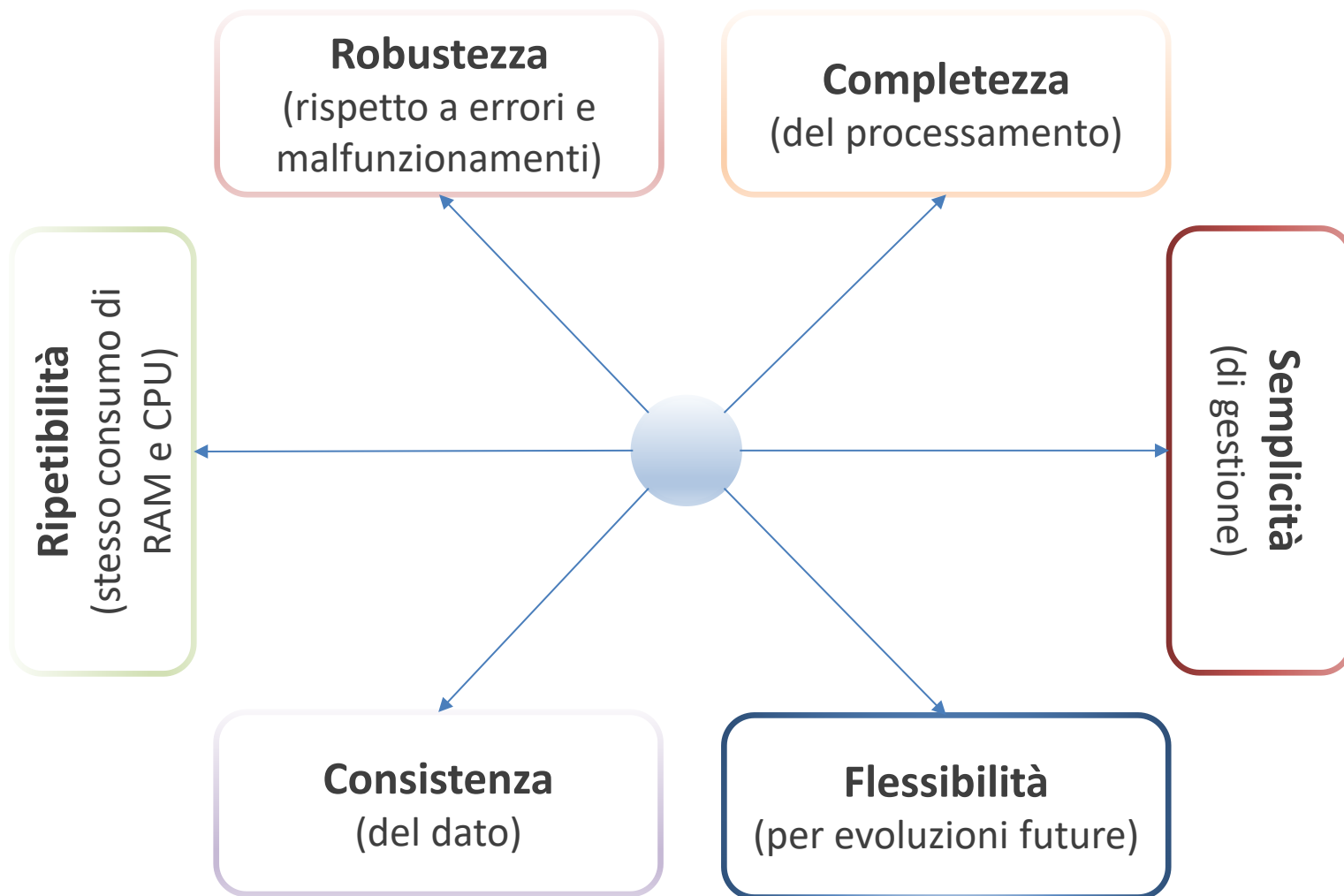
Load

Si carica il target a partire dalle Mirror Tables.

Precedenza stabilita dalle regole di integrità referenziale



Caratteristiche attese





Caratteristiche attese

Robustezza

Grazie al controllo di ripartenza, il riavvio dopo interruzione anomala può essere demandato a personale di control room, senza necessità di coinvolgere esperti della procedura

Ripetibilità

Sulla base della coordinata di flusso, si calibra l'esecuzione su di un quantitativo pressoché costante di dati. Con ciò, le risorse HW potranno rimanere costanti per lungo tempo



Caratteristiche attese

Consistenza

L'ambiente target deve assolutamente contenere dati consistenti. Cioè, non devono esistere record orfani (i.e.: con chiavi esterne inesistenti)

Ciò è garantito dal disegno. In caso di AbEnd:

- Precedente alla fase di "Load": nessuna scrittura è ancora avvenuta;
- Durante la fase di "Load": grazie al rispetto delle precedenze di caricamento già esposte;

Completezza

Riferita ai dati sorgente. La corretta individuazione della "coordinata di flusso", consente di orchestrare lo "scatto foto" in continuità: ciascun lancio parte dal record immediatamente successivo all'ultimo prelevato dal lancio precedente



Caratteristiche attese

Flessibilità

Lo sviluppo modulare così organizzato, corredato dal sistema dei "check point", consente l'apporto di interventi evolutivi con ridotta complessità di integrazione

Semplicità

La procedura così concepita incapsula la complessità realizzativa. Ne consegue che la gestione di regime può essere tranquillamente affidata a personale privo di conoscenze specifiche



Conclusioni

Evitando il più possibile di sviluppare "pezzi d'arte", favorendo lo sviluppo di un "metodo", si traduce in:

Consegne più controllabili

Velocità di apprendimento da parte di nuove risorse

Velocità di intervento in caso di segnalazioni del cliente

Valorizzazione dell'estro creativo dello sviluppatore

Il "metodo" diventa un "asset" aziendale, da sviluppare nel tempo



Per approfondimenti

Ralph Kimball, Margy Ross: The Data Warehouse Toolkit

-

-



GRAZIE PER L'ATTENZIONE

ing. Claudio Sanft

Claudio.sanft@delta-lambda.com

+39 348 320 2837