

ARCHITETTURA TECNICA

Sottosistema di Data Warehousing (DW)

INDICE

1	CONTESTO DI RIFERIMENTO ED OBIETTIVI	3
2	REQUISITI FUNZIONALI E TECNOLOGICI DELLA PIATTAFORMA DEL DW	4
3	MODELLO ARCHITETTURALE DI RIFERIMENTO.....	7
3.1	INTRODUZIONE	7
3.2	MODELLO CONCETTUALE	8
3.3	SCENARI D'EVOLUZIONE	10
3.4	COMPONENTI.....	11
3.4.1	<i>Staging Area</i>	11
3.4.2	<i>Enterprise Data Warehouse</i>	13
3.4.3	<i>Data Mart</i>	14
3.4.4	<i>Strumenti di Information Delivery</i>	15
3.4.4.1	Livello di reporting.....	18
3.4.4.2	Tipologie di strumenti di delivery	20
3.4.5	<i>Metadati</i>	22
3.4.5.1	Premessa	22
3.4.5.2	Tipologie di MetaDati	23
3.4.6	<i>Masterdata</i>	24
3.4.7	<i>Procedure ETL</i>	26
3.4.7.1	Acquisizione Dati.....	28
3.4.7.2	Caricamento dell'Enterprise Data Warehouse.....	30
3.4.7.3	Caricamento dei Data Marts.....	32
3.4.8	<i>Integrazione</i>	35
4	ARCHITETTURA TECNICA NEGLI SCENARI EVOLUTIVI IDENTIFICATI	37
4.1	ARCHITETTURA TECNICA	37
4.2	ARCHITETTURA OPERATIVA.....	38
4.2.1	<i>Procedure di Ritenzione dei dati</i>	38
4.3	ARCHITETTURA DEL MODELLO DATI.....	39
5	APPENDICI	41
	APPENDICE A – FLUSSO LOGICO DEI DATI	41
	APPENDICE B – DATA WAREHOUSE, MODELLI CONCETTUALI DI RIFERIMENTO	42
	APPENDICE C – METADATI PER AREE DEL DATA WAREHOUSE.....	45

1 CONTESTO DI RIFERIMENTO ED OBIETTIVI

Il disegno del SIS-N, ha previsto la presenza di una componente infrastrutturale di Data Warehouse che, grazie alla predisposizione di uno opportuno framework/linee guida di sviluppo, ha definito l'insieme delle corrette metodiche attraverso cui controllare e certificare i diversi momenti procedurali nel quale passa la vita del dato del SIS-N:

- **acquisizione del dato;**
- **elaborazione del dato;**
- **delivery dell'informazione.**

Tale componente gioca il ruolo fondamentale per integrare, anche concettualmente, le informazioni sparse su più basi dati operazionali e residenti su piattaforme logico-fisiche distinte, eliminando la possibilità che diversi fruitori del SIS-N giungano a risultati profondamente differenti pur osservando fenomeni comuni. Garantisce inoltre al procedimento analitico la correttezza e la completezza dei dati in esso contenuti.

Nel seguito del documento sono descritti:

- l'infrastruttura della piattaforma di Data Warehouse realizzata per SIS-N;
- l'insieme di tutti i principi e le linee guida adottate nell'ambito dei diversi progetti implementativi.

Nel seguito ci soffermeremo nella descrizione dei seguenti elementi:

- l'insieme complessivo dei macro requisiti tecnologici e funzionali che un generico sistema di DW deve offrire;
- il framework tecnologico ed applicativo implementato per SIS-N;
- la natura dei processi, dei dati e dei metadati gestiti all'interno delle diverse componenti dell'architettura;
- l'infrastruttura tecnica, prodotti e SW di base del SIS-N;

Nel rispetto delle attuali e future esigenze, la definizione dell'infrastruttura del data warehouse garantisce scalabilità ed estendibilità.

2 REQUISITI FUNZIONALI E TECNOLOGICI DELLA PIATTAFORMA DEL DW

In generale un sottosistema di Data Warehousing è costituito da metodi, tecnologie e strumenti di ausilio al “lavoratore della conoscenza” (sia esso dirigente, amministratore o analista) per condurre analisi finalizzate all’attuazione di processi decisionali e al miglioramento del patrimonio informativo.

Una delle caratteristiche fondamentali di un’architettura di Data Warehouse è quella di risultare come un insieme di processi e di tecnologie finalizzate a facilitare l’accesso e la fruizione del patrimonio informativo.

A differenza dei tipici sistemi informatici gestionali, un’infrastruttura di Data Warehouse deve essere concepita in maniera da poter rispondere in modo efficace ed efficiente alle diverse necessità degli utenti. Non può essere concepita come un sistema “finito” ma come un **processo in continua evoluzione**. Tra le affermazioni ricorrenti in letteratura si sottolinea che “il Data Warehousing non si compra, si costruisce”.

L’infrastruttura di Data Warehouse costituisce l’elemento che integra i dati prodotti all’interno di un’organizzazione con i dati acquisiti da fonti esterne.

Il Data Warehousing si propone di ridefinire l’architettura informativa attraverso un processo di selezione, di trasformazione, di consolidamento e di presentazione dell’insieme dei dati di qualità certificata, provenienti sia dall’interno che dall’esterno del sistema in cui esso si colloca.

Le modalità di attuazione dipendono ovviamente dal contesto applicativo e questo rende il processo di Data Warehousing sempre vario nei suoi aspetti implementativi; nel seguito si tratterà il caso specifico della sua attualizzazione all’interno del Ministero della Salute. Il modello complessivo, mostra come l’utilizzo di processi di Data Warehousing contribuisca a svincolare le funzioni di fruizione, per quanto possibile, dai restanti sottosistemi operazionali e da altre fonti esterne al sistema.

All’interno di tale sistema gli utenti possono trovare una fonte “certificata” in cui tutte le informazioni presenti all’interno dei vari sottosistemi informativi sono state depositate al fine di permetterne la fruizione a vari livelli.

I principi architetturali fondamentali per un sistema di Data Warehousing caratterizzanti pertanto anche il sottosistema del SIS-N, sono:

- *Separazione*: l’elaborazione analitica e quella transazionale sono disaccoppiate.
- *Disponibilità*: Il sottosistema nazionale DW non costituisce un componente talmente critico per le attività istituzionali del Ministero, da richiedere una disponibilità 7x24”. Tuttavia, il livello di disponibilità può essere differente per ognuna delle diverse componenti dell’architettura. In particolare, la componente maggiormente critica è rappresentata dal sottosistema dedicato alla ricezione ed al trattamento dei flussi informativi acquisiti attraverso il Sistema di Cooperazione. Tale sottosistema ha un alto livello di affidabilità e disponibilità per ricevere i flussi alimentanti senza che una loro prolungata latenza nel Sistema di Cooperazione determini una saturazione dello storage ad esso dedicato.
- *Scalabilità ed estendibilità*: l’architettura hardware e software è concepita in modo da poter essere adattata a fronte della crescita nel tempo dei volumi trattati e del numero di utenti.

Inoltre il sistema è estendibile, cioè aperto a nuove applicazioni e tecnologie, senza che ciò comporti la riprogettazione integrale del sistema o delle sue componenti. Un'architettura modulare garantisce una scalabilità ed estendibilità ottenuta in ogni area mediante logiche/soluzioni diversificate.

La caratteristica è rilevante per il sistema DW del Ministero, in particolare per quanto concerne la mole di dati da trattare in continua crescita. Inizialmente, la numerosità dei dati e le tipologie delle applicazioni erano limitate e quindi non hanno richiesto strumenti eccessivamente sofisticati sotto il profilo funzionale e tecnologico. Nel tempo si è verificata però una continua crescita nei volumi dei dati e nel numero delle applicazioni da realizzare. Ciò è coerente con la natura di ogni sistema di Data Warehouse (in continua evoluzione). Ne consegue che tali considerazioni hanno indotto alla seguente conclusione circa i criteri da adottare per la selezione dei componenti HW/SW di base (DBMS, ETL, ecc) che dovranno evolvere a fronte delle mutate esigenze.

- *Facilità di accesso ai dati:* Il sistema maschera all'utente tanto l'eterogeneità delle fonti alimentati quanto la complessità dei dati trattati, fornendo strumenti adeguati alle esigenze di utenti che operano tipicamente in modalità on-line. L'infrastruttura fornisce strumenti all'utente che gli consentano di interrogare la base dati senza conoscere la struttura fisica o linguaggi di programmazione (SQL).
- *Flessibilità nella fruizione dei dati:* Nell'infrastruttura sono disponibili diverse modalità di accesso alle informazioni (report tabellari e/o grafici predefiniti, strumenti per la formulazione dinamiche di Query, strumenti per la formulazione di analisi OLAP, strumenti per la selezione e l'estrazione di dati, ecc). Le informazioni sono consultabili ed utilizzabili attraverso i prodotti software attualmente a disposizione degli utenti finali (es: attraverso browser, etc).
- *Persistenza:* Le informazioni sono recuperabili e ricostruibili a seguito di eventuali perdite. Ciò non implica il mero backup-recovering, ma anche il rispetto del principio secondo il quale nessun dato deve essere perso nei processi di trasformazione cui è sottoposto. Le procedure d'elaborazione e trasformazione prevedono un opportuno meccanismo di tracciatura per ricostruire il ciclo di vita del dato. Tutti i dati acquisiti dal sistema sono sottoposti a procedure di salvataggio.
- *Qualità dei dati:* L'affidabilità dei dati resi pubblici in un Data Warehouse è cruciale. Per assicurare un elevato livello di affidabilità dei dati è necessario non solo filtrare i dati in ingresso per bonificarli, ma anche organizzare l'insieme dei processi previsti (estrazione, trasformazione, aggregazione, alimentazione) in modo da assicurare la corretta sequenza del flusso delle informazioni.
- *Sicurezza e profilatura:* Tutte le componenti dell'infrastruttura sono integrate all'interno dei sottosistemi di sicurezza e di profilatura già definiti in ambito SIS-N secondo le modalità descritte nei capitoli successivi, in aderenza alla normativa vigente.
- *Monitoring:* Nell'architettura di DW, è predisposta un'infrastruttura che permetta:
 - Il controllo ed il monitoraggio tecnologico delle diverse componenti dell'architettura al fine di poter rilevare, misurare, analizzare e prevenire problemi di funzionamento del sistema stesso (Performances e Tuning);
 - Il controllo ed il monitoring dei diversi flussi e dei relativi processi di elaborazione/trasformazione fornendo alle diverse classi di utenza coinvolte nell'intero processo di DW strumenti per poter verificare e controllare l'esito delle diverse elaborazioni e/o controllare la qualità/validità dei dati forniti.

Come si evince, la complessità intrinseca in un sistema di questo tipo è fondamentalemente rappresentata dalla necessità di conciliare tre aspetti palesemente contrastanti:

- Essere in grado di gestire grosse quantità d'informazioni;
- Garantire flessibilità agli utenti nell'accesso alle informazioni;
- Assicurare adeguati livelli prestazionali.

Il successo di un Data Warehouse è legato al raggiungimento di tali obiettivi primari per il cui soddisfacimento occorre conciliare capacità progettuali all'interno di un'infrastruttura HW/SW di base opportunamente definita. La soluzione logica implementata, per l'infrastruttura del sottosistema DW del Ministero della Salute, si basa su un'architettura in grado di:

- Assicurare la disponibilità di una fonte contenente la totalità delle informazioni raccolte sul territorio presso le entità coinvolte;
- Assicurare la coerenza e l'attualità delle informazioni presenti nel sistema;
- Prevedere un meccanismo di alimentazione che risulti veloce ed affidabile;
- Assicurare un adeguato livello prestazionale permettendo agli utenti del sistema di poter interagire in modalità On-Line.

Le soluzioni tecniche adottate per poter soddisfare tali obiettivi sono:

- La definizione di un modello dati a più livelli per arrivare ad informazioni organizzate indipendentemente dai processi sorgenti (modello a soggetti - vedi Modello Concettuale);
- La realizzazione di appositi processi di alimentazione del DW comprensivo di fasi di normalizzazione e di consolidamento dei dati;
- La realizzazione degli standard di alimentazione del DW, prevedendo la presenza di meccanismi in grado di controllare l'esito dell'intero processo in maniera automatica;
- La presenza di diversi livelli di rappresentazione delle informazioni, in particolare di un livello definito "di base" in cui i dati sono presenti al minimo livello di dettaglio e da più livelli aggregati. Questa strutturazione a "livelli" dei dati sarà presente in tutte le strutture logiche definite sia a livello di Data Warehouse sia a livello di Data Mart;
- La disponibilità di strumenti per supportare l'accesso alle informazioni da parte degli utenti finali

3 MODELLO ARCHITETTURALE DI RIFERIMENTO

3.1 Introduzione

Nel modello architetturale del “Sistema Informativo Sanitario Nazionale” (SIS-N) il data warehouse si integra con:

Sistemi sorgenti:

- Tutte le possibili fonti d'alimentazione del SIS-N. Le fonti possono essere esterne (Regioni, ASL, ...) ed interne allo stesso SIS-N.

Sistemi d'Infrastruttura:

- *Infrastrutture di Sicurezza* – Costituita dai sistemi d'autenticazione ed identificazione
- *Piattaforma di Accesso ai Servizi* - Consente l'accesso tramite strumenti di fruizione di larga diffusione.
- *Infrastrutture di Cooperazione* – Consente l'acquisizione dei dati dai sistemi esterni.

Sistemi per i Servizi di base:

- *Servizi Trasversali*
- *Catalogo dei Dati*

3.2 Modello Concettuale

Nella figura seguente si riporta il modello concettuale adottato per l'implementazione del sottosistema di Data Warehousing, denominato DWIDA (Data Warehouse Information and Delivery Architecture).

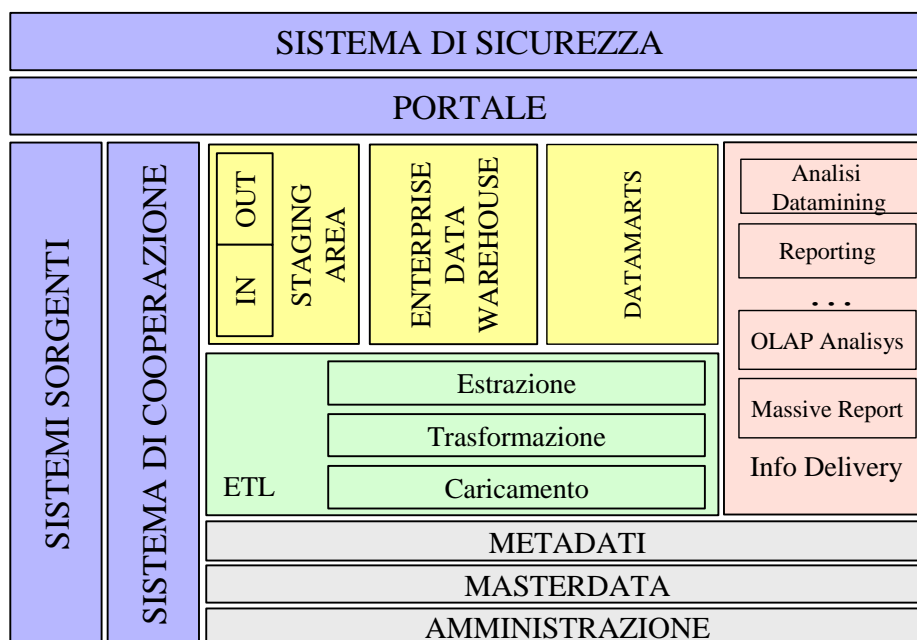


Figura 3.1 - Modello Concettuale

Si descrivono di seguito, sinteticamente, le componenti che costituiscono il modello.

- *Sistemi sorgenti* – costituiscono l'insieme dei sistemi che forniscono i dati d'input al sistema data warehouse.
- *Sistema di Cooperazione* – costituisce il canale di trasmissione dei dati all'interno del sistema SIS-N ed è quindi l'elemento alimentante il data warehouse.
- *Staging Area* – costituisce il database in cui saranno mantenuti i dati recuperati dal sistema di cooperazione. Durante l'acquisizione, i dati sono immagazzinati nel modulo di ricezione "IN". Sono quindi sottoposti a procedure per il controllo della qualità, eventuali segnalazioni d'errore e tutti i dati che verranno scartati, sono immagazzinati nel modulo "OUT". Tale modulo costituisce la fonte per analisi specifiche e per alimentare il sistema di cooperazione nel caso in cui le informazioni devono transitare verso sistemi esterni. Successivamente i dati che superano il processo di qualità, vengono consolidati per il trasferimento all'Enterprise Data Warehouse.
- *Enterprise Data Warehouse* – costituisce l'insieme dei dati di massimo dettaglio sulle molteplici tipologie d'informazione. Tali dati sono ottenuti attraverso successive operazioni di trasformazione ed elaborazione delle informazioni messe a disposizione nella staging area.
- *Data Mart* – Costituisce i database dove i dati saranno organizzati in modo da fornire una visione orientata ai soggetti d'interesse. La base dati informativa risultante è fortemente integrata, consistente e rappresentativa dell'evoluzione temporale dei fenomeni in essa memorizzati.

- *Prodotto ETL* – E' lo strumento di supporto alle attività di movimentazione dei dati nei tempi e nelle modalità predefinite (per esempio trasporto dei dati, gestione delle eccezioni, audits, ...).
- *Modulo di Amministrazione* – costituisce l'insieme delle procedure che permetteranno di gestire processi di ritenzione dei dati e storicizzazione.
- *Masterdata* o anche anagrafiche – costituiscono lo strato d'informazioni a supporto dell'information delivery, dell'arricchimento ed in alcuni casi della valorizzazione degli eventi elementari.
- *Metadata* – costituiscono strumenti e tecniche per la definizione, la raccolta e la pubblicazione delle informazioni relative alla strutturazione dei dati ("dati sui dati"), al significato di "business" (regole di business, indicatori di performance, ...) ed alle caratteristiche tecniche dei dati (formato, tempi di caricamento, ...)
- *Sistemi d'Info delivery, Portale e Sistema di Sicurezza* – permettono agli utenti un accesso personalizzato alle informazioni d'interesse.

3.3 Scenari d'evoluzione

Il modello concettuale è stato calato in un'architettura reale (hw, prodotti e strumenti software, ...) in grado di evolversi di pari passo con l'evoluzione delle esigenze del SIS-N.

Il sistema DW è stato progettato e realizzato tenendo conto dei seguenti elementi, che costituiscono aspetti critici rispetto alle possibili evoluzioni del sistema stesso.

- **Volume dei dati da elaborare**
 - Numero dei sistemi sorgenti
 - Numero dei flussi dati provenienti da ogni singolo sistema sorgente
 - Numero dei record per flusso in funzione del tempo
 - Grandezza media dei record
 - Frequenza di acquisizione/elaborazione dei dati
 - Tempo disponibile per l'esecuzione dei processi ETL
- **Volume dei dati da immagazzinare nel data warehouse**
 - Numero dei sistemi sorgenti
 - Numero dei flussi dati provenienti da ogni singolo sistema sorgente
 - Numero dei record per flusso
 - Grandezza media dei record
 - Modalità di retention e storicizzazione dei dati in funzione del tempo
- **Popolazione di utenti**
- **Modalità di fruizione** (es: datamart, reporting)

Ogni componente è stata acquisita/realizzata, "premiando" la salvaguardia degli interventi in caso di modifica delle altre.

- I moduli di Staging Area, EDW, Data Marts, Metadati e Masterdata, sono fisicamente costituiti da basi dati. Per permettere la scalabilità e l'adattabilità del sistema, l'analisi tecnico/funzionale si è basata su tecnologie di processi ed elementi logici standard per tutte le piattaforme.
- L'implementazione di un data warehouse in un contesto in fase di evoluzione come quello SIS-N, richiede nel tempo la necessità di trattare nuove tipologie d'informazioni, di interfacciarsi con periferiche sorgenti differenti da quelle attuali, e di gestire nuovi modelli dati per il delivery. Tali necessità si traducono tecnicamente nell'utilizzo di strutture hardware evolute, il cui disegno potrà evolvere nella complessità e nella numerosità degli elementi. Ad esempio l'aumento delle informazioni provenienti dai sistemi sorgenti, possono richiedere:
 - capacità di calcolo superiore o una distribuzione dei processi su unità di calcolo differenti (nuove CPU oppure integrazioni di altri nodi/server).
 - aumento degli storage dei database.
 - aumento della memoria fisica (RAM) .
 - potenziamento delle reti intranet.

A tale proposito, anche la scelta dei prodotti software utilizzati (sistema operativo, database, ETL, ec.) è stata operata in modo da garantire la gestione di tali evoluzioni, in termini di tecniche di clustering, di parallelizzazione dei processi, ecc.

-

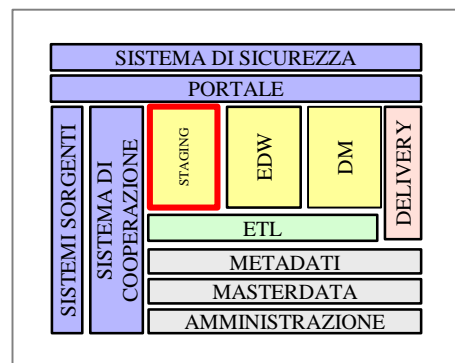
3.4 Componenti

3.4.1 Staging Area

INPUT: Sistema di Cooperazione

OUTPUT:

I dati consolidati costituiscono la fonte d'alimentazione dell'EDW. Per quanto riguarda analisi specifiche sui dati pervenuti dai sistemi sorgenti, la staging area alimenta direttamente i prodotti di delivery. Inoltre le informazioni scartate ed eventuali segnalazioni sono messe a disposizione ai sistemi esterni interessati attraverso il sistema di cooperazione.



FUNZIONE:

L'obiettivo primario della Staging Area è quello di disaccoppiare i sistemi sorgenti ed il sottosistema nazionale DW. Tutti i flussi informativi per poter essere elaborati sono inizialmente caricati all'interno della Staging Area ove sono soggetti a processi di validazione, trasformazione e validazione. La componente infrastrutturale HW/SW di base che ospita questo componente è predisposta per permettere una scalabilità idonea a garantire un livello di disponibilità tale da assicurare l'elaborazione dei flussi senza creare criticità al sistema di cooperazione applicativa che costituisce di fatto il canale di riferimento attraverso cui pervengono i flussi.

La fase di estrazione è concepita in modo da ridurre al minimo indispensabile il tempo di latenza dei flussi sul sistema di cooperazione, quindi essa è fondamentalmente caratterizzata da un insieme di processi che si preoccupano essenzialmente di effettuare lo spostamento dei dati dal sistema di cooperazione alla staging area. In questa fase non vengono effettuati controlli di nessun tipo.

Quella di Staging è **un'area di transito che non ha responsabilità analitiche sui dati**; essa è sufficientemente strutturata da permettere un efficace trattamento dei dati prima del loro passaggio ai livelli superiori.

I processi di ETL previsti in ambito Staging Area sono stati predisposti in modo da prevedere la gestione dei flussi di estrazione (dal sistema di cooperazione), caricamento (in staging), validazione (in staging), trascodifica (in staging) e consolidamento (in staging).

Al termine della fase di consolidamento e successiva alimentazione dell'EDW, i dati non hanno più ragione di permanenza nella Staging Area se non per eventuali motivi tecnici. Ad esempio, nel caso in cui i sistemi alimentanti forniscono snapshot periodici totali delle anagrafiche e dei dati, viene mantenuta nell'Area di Staging l'ultima immagine in modo da consentire il calcolo dei delta ed eseguire i successivi processi di trasformazione utilizzando tali strutture come elementi di raccordo. Dettagli ulteriori sulla tipologia di strutture dati che potranno essere ospitate in questa area saranno esposti nel capitolo relativo alla gestione delle Anagrafiche/Masterdata.

Pur non essendovi, in generale, un periodo prefissato a priori di permanenza dei dati nella Staging Area, i flussi vi rimangono in latenza il tempo necessario per:

- Sincronizzarsi con eventuali altri flussi, fino a soddisfare tutte le condizioni formali per la propria elaborazione.
- Essere correttamente elaborati
- Permettere il recupero e la gestione degli scarti
- Calcolare i delta dove i flussi in ingresso siano degli snapshot

Una volta soddisfatte tutte queste condizioni, **un flusso correttamente elaborato è eliminato dalla Staging Area dopo l'eventuale migrazione su un apposito sistema di Backup**. Infatti tali dati, pur non essendo più utili ai fini del processo di DW, possono essere comunque archiviati ed i relativi riferimenti inseriti all'interno di un catalogo in modo da poter sempre consentire la ricostruzione della storia dei flussi ricevuti dai diversi sistemi sorgenti. La gestione del salvataggio dei flussi acquisiti è gestita da un componente infrastrutturale che mette a disposizione il servizio per i diversi progetti implementativi. È compito dello specifico progetto valutare (e giustificare) la necessità di una gestione dei backup dei file pervenuti al sistema per la loro elaborazione.

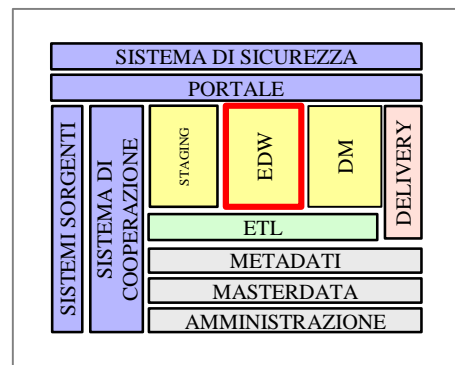
3.4.2 Enterprise Data Warehouse

INPUT: Staging Area

OUTPUT: Data Marts

FUNZIONE:

L'Enterprise Data Warehouse (nel seguito indicato come EDW) **costituisce l'infrastruttura dati fondamentale per il supporto dei flussi di distribuzione**. Il termine Enterprise descrive una tipologia di data warehouse orientata alla **gestione di molteplici tipologie d'informazione**; tipicamente un EDW è alimentato da differenti sorgenti. L'EDW presenta le seguenti caratteristiche:



- Ambiente di consuntivazione dei dati base (eventi elementari) su logiche relazionali.
- Elevata granularità del dato.
- Tecnologia relazionale con partizionamento spinto e disegno orientato alla gestione di grossi volumi.
- Costituisce il centro della distribuzione dei dati sulle strutture dedicate all'analisi.
- Vista l'estrema analiticità delle informazioni, raramente offre supporto ad applicazioni di delivery.
- La struttura logica è solitamente molto semplice, il disegno fisico bilancia esigenze di efficienza nello storage e nella distribuzione dei dati.
- Garantisce la completa tracciabilità e ricostruibilità delle informazioni in coerenza con quanto presente nelle viste end - user dei Data Marts.

Dal modello concettuale di riferimento precedentemente illustrato, si determina anche il processo di alimentazione dell'EDW. In particolare tutti i dati resi disponibili dalle interfacce alimentanti vengono mappati e gestiti attraverso operazioni di caricamento, certificazione, integrazione e altre trasformazioni specifiche per tipologia e natura del dato. Tutte le operazioni appena specificate sono ottenute mediante la parametrizzazione dei metadati su tabelle relazionali.

Successivamente il dato, attraverso determinate regole di aggregazione e trasformazione, viene riportato sui Data Marts che costituiranno le strutture "end user".

E' presente nell'EDW un ulteriore componente: l'Operational Data Store (ODS). **Esso è utilizzato per archiviare i dati operazionali di dettaglio** risultanti dal processo di integrazione e ripulitura dei dati sorgente in un formato normalizzato, prima che gli stessi siano aggregati nel Data Warehouse. Tale componente **permette di avere un modello di dati comune e di riferimento per l'intero sistema e di gestire separatamente le problematiche legate all'estrazione ed all'integrazione dei dati** da quelle riguardanti tipicamente l'alimentazione del DW.

L'ODS è aggiornato, tendenzialmente, in tempo reale, da sistemi esterni o direttamente dagli utenti. Questo archivio ha essenzialmente la funzione di soddisfare requisiti prestazionali, di scalabilità e reportistica in tempo reale.

3.4.3 Data Mart

INPUT: Enterprise Data Warehouse e Staging Area

OUTPUT: Sistemi di Delivery

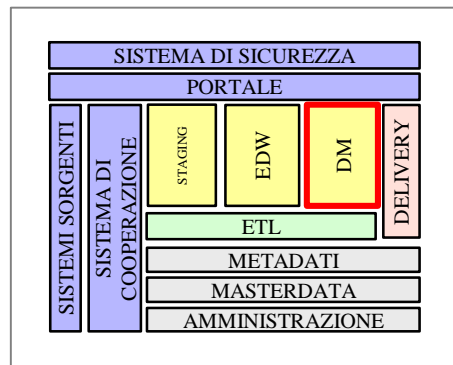
FUNZIONE:

I Data Marts sono un sottoinsieme proprio dell'EDW orientato tipicamente ad un solo settore aziendale.

In funzione delle esigenze di business intelligence e dell'ottimizzazione tecnica-applicativa **queste strutture hanno diversi gradi di aggregazione**, di raggruppamento e di indicizzazione. Per garantire la "consistenza" delle informazioni ai diversi livelli dell'architettura i datamart sono preferibilmente alimentati da un'unica fonte.

Le aree dei datamart sono definite in modo da garantire:

- L'aggregazione dei dati elementari in tabelle fisiche disegnate in modo funzionale alle analisi richieste. La determinazione della stessa tipologia di dati in modo di ridurre alla sola estrazione dei dati la fase di delivery dell'informazione nella reportistica.
- Un disegno tale da essere in grado di supportare differenti tipologie di reportistica, analisi etc (informazioni utili agli utenti). L'accesso alle informazioni in tempo reale oppure in modalità batch, con preformattazione di report e viste.
- Che la struttura di un data mart contenga le KPI (o indicatori) localizzati nelle strutture (focchi o stella).
- La possibilità di inserire e cancellare data mart senza alcun impatto sull'architettura.



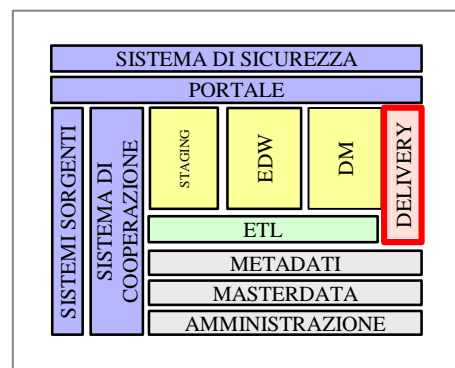
3.4.4 Strumenti di Information Delivery

INPUT: DataMarts, Enterprise Data Warehouse, Staging Area

OUTPUT: End-Users

FUNZIONE:

La componente di Delivery costituisce, nell'ambito dell'infrastruttura DW, l'**elemento architetturale preposto alla produzione, divulgazione e distribuzione delle informazioni** secondo logiche e metodiche diversificate in funzione delle classi di utenza del sistema. In generale, come evidenziato nel seguito, il canale di accesso preferenziale per l'accesso alle informazioni è costituito dal canale Web/internet. Attraverso esso si possono consultare report predefiniti, richiamare applicazioni di consultazioni custom, eseguire limitate operazioni di query e reporting. Dal punto di vista funzionale, la presenza di questo elemento architetturale nel sistema di Data Warehousing consente di soddisfare le seguenti tipologie di esigenze:



- **Query** - Interrogazione delle strutture dati per fini specifici o per l'estrazione d'informazioni o flussi strutturati. È la modalità d'interazione di più basso livello e richiede una specifica competenza sul dominio dei dati.
- **Reporting** - Realizzazione ed esecuzione di reportistica predefinita che struttura e standardizza l'accesso alle informazioni, anche nelle modalità e tempistiche (periodicità) di distribuzione.
- **Inquiry** - Possibilità di costruire viste e reportistica occasionale in un ambiente grafico di facile uso anche per chi non ha conoscenze di linguaggi d'interrogazione dati.
- **Analisi** - Accesso dinamico ai dati a partire da viste organizzate che consentano di evidenziare informazioni statistiche quali medie, andamenti, tendenze rispetto a particolari sottoinsiemi di dati. Tra questi strumenti rientrano quelli che permettono modalità di fruizione dati di tipo OLAP, ovvero che consentono l'analisi e l'esplorazione interattiva dei dati sulla base di modelli multidimensionali, aggregando ed esplodendo i dati mediante i comuni operatori di roll-up, drill-down, slice-and-dice, pivoting, drill-across e drill-through.
- **Estrazioni** - Possibilità di poter estrarre dal sottosistema di DW flussi da trasferire sui sistemi degli Enti richiedenti il dato per eventuali elaborazioni locali e la predisposizione di flussi contenenti segnalazioni relative alle elaborazioni dei dati ricevuti dai sistemi esterni. Come per l'ingresso dei dati, anche la relazione di uscita verso eventuali sistemi destinatari, avviene attraverso il sistema di cooperazione. Compito della piattaforma di Delivery è quello di produrre i flussi nel giusto formato e di depositarli nella porta di dominio di output.
- **Monitoraggio** - Possibilità di disporre di strumenti per il monitoraggio dell'intero processo di DW da parte degli amministratori del sistema (per funzionalità di tuning tecnologico dell'infrastruttura e per funzionalità di controllo funzionale mediante la verifica dello stato dei diversi processi di alimentazione).
- **Applicazioni Specifiche** - disponibilità di applicazioni di consultazione e navigazione dei contenuti informativi gestiti dal sistema mediante applicazioni guidate che forniscono un mezzo agli utenti per accedere in modo controllato ai dati secondo opportune logiche di business.
- **Knowledge Discovery** - Disponibilità di tecniche evolute di Data Mining a supporto di attività mirate all'individuazione di informazioni nascoste dalla mole dei dati memorizzata nel Data Warehouse e nei Data Marts, sia con fini *descrittivi* che *predittivi* dei fenomeni d'interesse.

Dall'elenco sopra riportato sono escluse le tipiche funzionalità di stampe massive che, nello scenario architetturale complessivo del SIS-N, sono di competenza della componente transazionale. Considerando, invece, le diverse tipologie di utenze, queste possono essere classificate in:

- **Utenti Amministratori Tecnici:** Sono costituiti da DBA, Sistemisti, ed in genere dagli specialisti dei prodotti. Sono responsabili della verifica e del corretto funzionamento delle componenti HW/SW dell'infrastruttura tecnica;
- **Utenti Amministratori Funzionali:** Sono costituiti da responsabili del controllo e del monitoraggio delle diverse componenti applicative (procedure di ETL, alimentazione dei DM, verifica della disponibilità delle applicazioni di reportistica, verifica del livello di qualità dei dati acquisiti, ecc).
- **Utenti interni al SIS-N:** costituisce la classe di utenza che utilizza l'insieme complessivo di tutti gli strumenti e delle applicazioni di delivery messe a disposizione dall'infrastruttura del sottosistema nazionale DW per l'accesso e la consultazione delle informazioni. Questa classe di utenza si può suddividere in:
 - **Utenti di base:** Costituisce la classe di utenza che necessita di applicazioni e strumenti di consultazione guidata.
 - **Utenti evoluti:** Costituisce la classe di utenza che oltre all'utilizzo di applicazioni e strumenti di consultazione guidata può anche utilizzare strumenti di query e reporting che permettono un certo livello di libertà nelle modalità di accesso ai dati.
 - **Utenti specializzati:** E' la classe di utenza che necessita di strumenti di analisi evoluti quali tool di Data Mining.
- **Utenti esterni al SIS-N:** classe di utenti che possono accedere ai servizi di delivery messi a disposizione dal sottosistema nazionale DW per eseguire consultazioni guidate ed, eventualmente, estrazione di flussi.

La distribuzione delle tipologie di funzionalità di delivery, sulle diverse classi di utenza sopra definite, è sintetizzata nella successiva tabella.

	Query	Reporting	Inquiry	Analisi	Estrazioni	Monitoraggio Tecnico	Monitoraggio Funzionale	Applicazioni Specifiche	Knowledge Discovery
Utenti Amm. Tecnici						X			
Utenti Amm. Funzionali SIS-N							X		
Utenti SIS-N di base		X	X		X			X	
Utenti SIS-N evoluti	X	X	X	X	X				
Utenti SIS-N specializzati	X	X	X	X	X				X
Utenti esterni a SIS-N		X	X		X			X	

Come evidenziato dalla figura seguente, la componente di Delivery opera su tutti i livelli in cui è strutturato il modello dati del sottosistema nazionale DW; in funzione della tipologia di funzionalità richiesta e dalla tipologia di utente.

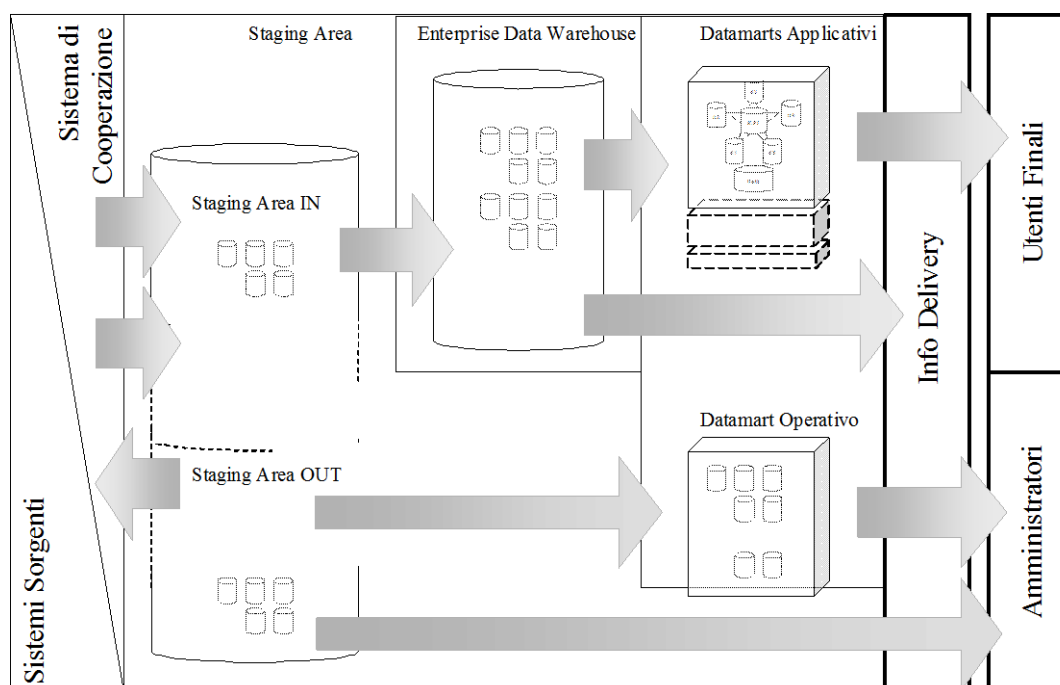


Figura 3.2 - Flusso dei dati per il delivery

3.4.4.1 Livello di reporting

Il modello architetturale utilizzato per il sottosistema di Data Warehousing, come descritto nelle precedenti sezioni, prevede i seguenti livelli di gestione dei dati:

1. Staging Area
2. Enterprise Data Warehouse
3. Data Marts
4. Metadati
5. MasterData

In generale:

- il livello di **Staging Area** ricopre essenzialmente il ruolo di disaccoppiare i sistemi sorgenti dall'EDW, pertanto essa svolge una funzione di natura tecnica. I dati che risiedono all'interno di questo strato del modello architetturale sono essenzialmente funzionali alle procedure di ETL; quindi non è prevista nessuna possibilità di rilascio di strumenti e funzionalità di delivery per gli utenti finali. Diverso è, invece, il discorso per quanto riguarda gli utenti amministratori che devono disporre di funzionalità specializzate per il controllo e la verifica dei processi di ETL e dei log applicativi di questi prodotti. Come descritto nella sessione relativa ai **Metadati**, tutte le informazioni generate dai processi di ETL (di tipo non solo tecnico ma anche funzionale/applicativo), sono organizzate in un apposito Data Mart di controllo trasversale a tutte le specifiche applicazioni su cui vengono costruite funzionalità specifiche per gli utenti amministratori centrali/regionali.
- Il livello preposto alle attività di Query & Reporting e di analisi è quello dei **Data Marts**, per cui ogni esigenza analitica di business può essere innanzitutto soddisfatta da un apposito o adeguato Data Mart specializzato per singola tematica applicativa/funzionale o di area. A questo/i fa riferimento lo strato semantico di disaccoppiamento (**Metadati**) e tutti i conseguenti reports di analisi. Queste strutture informative sono organizzate in modo da privilegiare le funzionalità di consultazione per cui lo schema logico di riferimento è costituito essenzialmente da "Star Schema" e da tabelle pre-aggregate integrate con l'insieme di tutti i dati elementari che hanno concorso alla formazione/elaborazione dei dati presenti nelle strutture sommarizzate. La scelta architetturale di includere in ogni Data Mart l'insieme complessivo di tutti i dati elementari relativi alla tematica trattata dallo stesso pone i seguenti vantaggi:
 - A livello di funzionalità, possono essere messe a disposizione anche applicazioni di analisi analitiche sui singoli eventi e/o la possibilità di attivare funzioni di drill dalle informazioni aggregate al singolo dettaglio;
 - A livello di scalabilità, includendo in ogni DM i relativi dettagli, si riduce il carico elaborativo sull'EDW. Nuovi Data Mart possono essere definiti nell'infrastruttura aumentando solo il carico dei processi di ETL sullo strato di EDW.

Infine, esistono particolari tipologie di Data Mart specificatamente orientati a fornire una risposta funzionale alle esigenze di applicazioni di Data Mining particolari che possono richiedere anche modelli di organizzazione e strutturazione dei dati diversi dal classico modello concettuale a "Star Schema" e che possono basarsi su modelli fisici di rappresentazione diversi da quello relazionale.

- Una limitata possibilità d'interrogazione è prevista sul livello **dell'Enterprise Data Warehouse**, ma non per soddisfare esigenze di business analitico. Le interrogazioni sul livello EDW hanno motivazioni di carattere gestionale, risponderanno a esigenze di reportistica non dimensionale o supportano eventuali estrazioni per sistemi esterni.
- Trasversali a questa distinzione di livelli si collocano tutte le strutture atte ad ospitare e gestire i **Metadati** e le anagrafiche del sistema (**MasterData**). Queste strutture, oltre che essere di supporto alle applicazioni/strumenti previsti per il delivery dell'informazione, possono essere considerate esse stesse delle fonti informative interrogabili e navigabili in particolare per:
 - Dare la possibilità agli utenti di valutare la natura, il livello di disponibilità e la tipologia delle informazioni presenti nel sistema;
 - Discernere tra le possibili dimensioni di analisi presenti nel sistema e le misure disponibili;
 - Valutare la rispondenza dell'informazione richiesta per le diverse funzionalità, in funzione delle regole di business utilizzate dal sistema di calcoli, allarmi, soglie e notifiche.

Le tipologie di funzionalità e di utenti coinvolti sono tali da richiedere necessariamente l'integrazione all'interno della componente di Delivery di più classi di strumenti integrati funzionalmente con una serie di applicazioni specifiche oltre che, dal punto di vista tecnologico, attraverso una modalità di condivisione dello strato di metadati. Tutti questi "canali" di erogazione dell'informazione sono inclusi all'interno di una componente di raccordo unica attraverso cui gli utenti possono accedere al sistema d'Analisi Direzionali e richiedere quanto di competenza secondo gli standard di sicurezza e nelle modalità definite a livello dell'intera architettura del SIS-N. Gli utenti, in funzione dello specifico livello di abilitazione/autorizzazione posseduto, possono:

- Accedere alle classi di report "certificati" predisposti a livello centrale;
- Effettuare operazioni di modifiche sulle strutture dei report predisposti centralmente;
- Accedere al proprio patrimonio informativo di competenza per eseguire analisi libere e/o guidate;
- Richiedere estrazioni ed invii periodici di flussi in funzione di specifiche policy di abbonamento/sottoscrizione;
- Accedere alle applicazioni custom predisposte dall'Amministrazione;
- Consultare cruscotti per valutare l'andamento di particolari indici di qualità (KPI);
- Consultare cruscotti di amministrazione per verificare lo stato di aggiornamento dei flussi e per verificare l'esito e la tipologia di eventuali segnalazioni di anomalie riscontrate dal sistema durante l'elaborazione;
- Accedere al patrimonio dei Metadati tecnici e funzionali

L'esigenza di disporre di funzionalità che consentano di gestire i profili di abilitazione e configurazione per utente/gruppo di utenti (ruolo) è soddisfatta grazie all'utilizzo, all'interno dell'infrastruttura, del componente Profile Manager all'interno del sottosistema di sicurezza.

Sono utilizzati appositi connettori software sviluppati ad hoc per alimentare i diversi sistemi di profilatura propri dei diversi strumenti utilizzati nella componente di delivery.

3.4.4.2 Tipologie di strumenti di delivery

Per coprire tutte le esigenze Query & Reporting, è necessario uno strumento che permetta un livello di crescente libertà nel trattamento dei dati, in relazione all'utenza. In particolare è garantita la copertura di tre specifici livelli di utilizzazione:

- **Livello 1** - Produzione e distribuzione di Report predefiniti centralmente sia in termini di struttura (contenuto) sia in termini di frequenza di elaborazione. Questa prima tipologia di esigenze, essendo costituita da Report costruiti centralmente e sui quali l'utente non può effettuare nessuna operazione di personalizzazione (se non l'impostazione dei parametri in ingresso) richiede la *certificazione sia a livello di contenuto informativo e sia dal punto di vista dei tempi di elaborazione attesi*. Questa specifica esigenza funzionale è erogata tramite canale Web.
- **Livello 2** - Introduzione di uno strumento per l'esecuzione di Report statici a livello di struttura (contenuto) ma parametrici a livello di filtri di selezione con la possibilità di effettuare il drill up/down, pivoting, ecc. In sintesi, i dati resi disponibili e i criteri di estrazione sono fissi, ma vi è un certo grado di libertà nell'impostazione del layout di presentazione. Analogamente a quanto previsto per gli oggetti del livello 1, *anche tale esigenza è certificata sia dal punto di vista del contenuto informativo e sia dal punto di vista dei tempi di elaborazione attesi*. Questa specifica esigenza funzionale è erogata tramite canale Web.
- **Livello 3** - Introduzione di uno strumento per la creazione dinamica di Report sia in termini di struttura (contenuto) sia in termini di filtri di selezione da applicare sui dati. Tale esigenza prevede sostanzialmente che l'utente (numero contenuto di esperti) possa eseguire analisi libere sui dati a partire da una loro strutturazione logica che riduca la complessità delle strutture fisiche. Per questo tipo di esigenza non è possibile assicurare gli stessi livelli di affidabilità in precedenza descritti; pertanto, essa è limitata a poche e certificate classi di utenza specialistica che dovranno disporre anche delle piattaforme client adeguatamente predisposte per poter eseguire in locale funzionalità avanzate di query & reporting, eventualmente rese disponibili anche attraverso il canale Web.

Questi tre livelli, pur essendo potenzialmente orientati a soddisfare esigenze di diverse classi di utenza (vedi tabella associazione utente/funzione), sono funzionalmente coperti da un unico strumento in modo da garantire il massimo livello di integrazione. Le caratteristiche tecniche che lo strumento di delivery soddisfa sono:

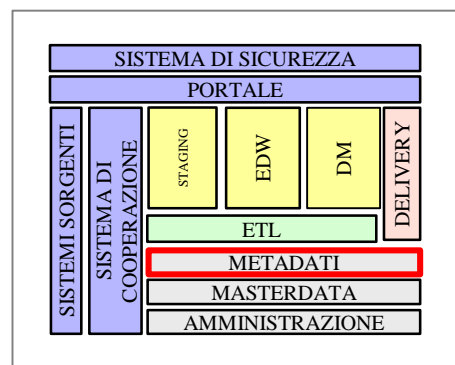
- produzione e pubblicazione di reports e templates in modalità Web Based, con logiche di categorizzazione e classificazione degli stessi ;
- creazione di reports altamente sofisticati con un numero elevato di dimensioni di analisi e con gestione di formati multipli (tabellare, grafico, documenti di testo, etc.);
- accesso a fonti dati esterne tramite, ad esempio ODBC, XML;
- esportazione di tutti i reports prodotti nei più comuni formati di file, quali HTML, XML, TXT e PDF, per integrare le analisi dettagliate con altri tools per applicativi desktop (ex. Excel);
- possibilità di automatizzare e schedare l'esecuzione dei reports e il rinfresco dei dati, anche al fine di aumentare il livello di efficienza e di scalabilità del sistema;
- accesso e manipolazione di reports (modelli di analisi) generati nell'ambiente di sviluppo;
- possibilità di invio dei reports prodotti ad altri utenti/gruppi di utenti mediante un sistema e-mail conforme allo standard MAPI;

- possibilità di definire regole di alert e meccanismi di notifica (ex. mail o invio di reports a gruppi/utenti specifici) al verificarsi di determinate condizioni;
- possibilità di impostare regole di calcolo per informazioni aggiuntive e derivate;
- navigazione multidimensionale;
- disponibilità di tecniche di *report bursting*, attraverso cui è possibile produrre un unico report e poi permettere l'accesso all'utente ai soli dati di competenza in modo da limitare il numero di oggetti da produrre;
- possibilità di sottomissione di query ad-hoc;
- accesso al Repository dei metadati sia in modalità C/S sia in modalità Web;
- modalità di visualizzazione e stampa "WYSIWYG".

3.4.5 Metadati

3.4.5.1 Premessa

I metadati costituiscono un substrato informativo necessario per la gestione di un sistema complesso e d'ausilio per fornire valore aggiunto alla soluzione. La loro valenza non è limitata al supporto tecnico ed operativo per il trattamento e la gestione dei dati, ma porta un valore meta-informativo utile per garantire visibilità, tracciabilità e documentabilità dei modelli e degli applicativi sviluppati nell'ambito dell'infrastruttura di DW del SIS-N.



Le meta informazioni che devono essere gestite non si limitano ai soli metadati cosiddetti Tecnici, vale a dire tutti quei dati finalizzati alla gestione puramente operativa dell'informazione, alla parametrizzazione ed alla sincronizzazione dei processi e delle procedure di trasporto dei dati, ma riguardano anche e soprattutto all'insieme complessivo di tutte le informazioni (siano esse regole e/o valori di riferimento) che stabiliscono regole atte a:

- validare la correttezza funzionale del dato e la sua eventuale completezza;
- determinarne le regole con cui l'informazione deve essere fruita e/o aggregata;

In tale contesto, un ruolo rilevante è rappresentato dall'insieme dei metadati preposti a verificare e misurare la qualità dell'informazione (*KPI di qualità Check*). La determinazione puntuale di tali KPI costituisce un elemento cruciale al fine di garantire la corretta alimentazione dei contenuti informativi. La loro determinazione costituisce uno dei primi obiettivi preliminari e/o paralleli all'attuazione dei vari progetti implementativi, in quanto costituiranno gli elementi di controllo a cui tutte le procedure di ETL dovranno fare riferimento. In un'ipotesi d'approccio incrementale alla determinazione/definizione dei metadati (approccio che appare il più probabile nell'attuale contesto), l'architettura implementata si pone l'obiettivo di fornire una visione unitaria dei metadati attraverso la realizzazione incrementale dei seguenti modelli di riferimento:

- **CBM: Common Business Model.** Il cui principio guida è che ogni dato censito deve essere definito secondo regole di standard definite, condivise e documentate in modo da mettere a disposizione un dizionario di riferimento comune a tutte le applicazioni.
- **CBR: Common Business Rules.** Analogamente a quanto detto per l'aspetto dati, anche per la definizione delle regole di validità/trasformazione/trattamento di essi, è necessario che siano seguiti gli stessi principi guida. La costituzione di un unico repository di metadati che raccolga le regole di validità dei singoli dati, porta un valore che si accresce progressivamente con la realizzazione di nuove applicazioni e porta a regole di integrità comuni ed esplicite, presupposto imprescindibile per garantire la qualità delle applicazioni future e la correttezza del valore dell'informazione.
- **CDM: Common Dimensional Model.** È il 'Dimensional Bus' definito da Kimball, che tende ad avere un'unica struttura dimensionale trasversale alla molteplicità dei Data Marts. Nel contesto del SIS-N ciò significa che è stato definito in modo puntuale l'insieme complessivo di tutte le componenti Anagrafiche principali (Master Data), come illustrato nel relativo capitolo.

Tali modelli, da un punto di vista tecnologico, sono costruiti a partire da una federazione di repository che mette in relazione i singoli repository di prodotto con un repository centrale per le informazione di ponte ed aggiuntive.

3.4.5.2 Tipologie di MetaDati

Metadati Tecnici

Come accennato, **la funzione principale dei metadati tecnici è di assicurare la congruenza delle informazioni caricate nel Data Warehouse con quelle dei sistemi sorgenti**, con l'obiettivo preciso di eliminare tutte le soglie di indeterminazione sullo stato dei dati e permettendo in qualsiasi momento di chiarire quale fase di processo stia attraversando un flusso alimentante, con quali esisti e quali scarti. Il principio fondamentale che nessun dato deve andare perduto non significa che dovrebbero esserci solo dati assolutamente corretti; significa piuttosto che deve essere possibile determinare con esattezza lo stato dei dati e ricostruire in ogni momento una situazione coerente. Vi sono quindi metadati per la descrizione del processo di acquisizione, controllo e trasformazione dei dati e sono ospitati su un repository comune al quale faranno costantemente riferimento i vari processi dell'ETL (ai tre livelli) per reperire informazioni e per depositarne altre. Avere questo tipo di dati su un repository comune non facilita solo la gestione ed il controllo dei processi ma permette anche di documentarli nella loro struttura, durante la loro esecuzione e per le caratteristiche prestazionali.

Metadati di "Business"

Classifichiamo come metadati di Business tutti quelli orientati alla fruizione del dato, al suo controllo semantico e alla verifica funzionale dello stesso.

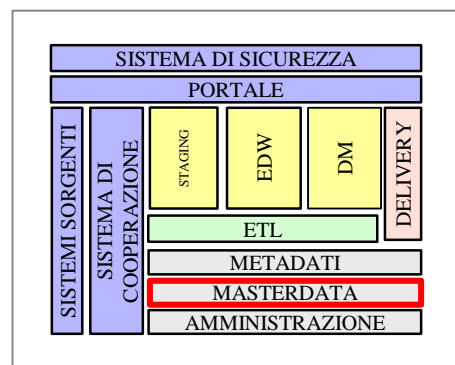
Molti metadati di questo tipo costituiscono regole che è bene portare a fattore definendo criteri semantici legati alle specifiche di business che, incrementalmente, attraverso i singoli progetti sviluppati, producano un modello di business condiviso ed un patrimonio informativo di valore.

Dalla chiarezza sulle regole di business discende anche una facile ed efficace gestione e strutturazione dei metadati tecnici che ne vadano a supportare la gestione operativa.

3.4.6 Masterdata

Verranno nel seguito affrontate le problematiche di gestione delle anagrafiche in relazione alla struttura a più livelli descritta, secondo le specificità proprie di ciascun contesto; il discorso si riferisce in particolare alle Anagrafiche ma alcune considerazioni di carattere più generale valgono anche per le altre tipologie di flusso.

Da un punto di vista architetturale, **l'insieme complessivo delle anagrafiche principali dell'intero sistema risiedono in un apposito strato del modello, rappresentato negli schemi architetturali proposti con il nome di MASTER DATA.**



Tali Anagrafiche sono opportunamente rese disponibili all'interno del livello di Masterdata attraverso opportuni processi di replicazione che provvederanno, su base periodica, ad allineare i diversi dati.

Le procedure di ETL preposte all'aggiornamento delle anagrafiche costituiscono un insieme specializzato di processi. Infatti, in funzione delle specifiche anagrafiche gestite si preoccupano di:

- tenere traccia delle variazioni;
- individuare le informazioni variate ed aggiornare le relative strutture compresi i metadati associati;
- gestire la storicità delle anagrafiche.

Oltre alla presenza di Anagrafiche gestite ed aggiornate esclusivamente centralmente, esistono Anagrafiche strutturate in modo da prevedere una fase di consolidamento di tipo "incrementale". Questo può essere il caso in cui i flussi in ingresso a livello nazionale e successivamente trasportati nella Staging Area provengono da fonti diverse con aree di sovrapposizione per quanto riguarda:

- dati di dominio
- dati di anagrafica

La sovrapposizione può essere duplice:

- da più fonti arrivano domini/anagrafiche sovrapposte (ad esempio la stessa tabella di dominio sulle possibili prescrizioni mediche)
- da fonti diverse arrivano dati disgiunti ma in relazione alla stessa entità logica (ad esempio, ogni regione/provincia invia i dati con formati e contenuti a livello di attributi e domini, diversi per singolo flusso)

Questi primi ma importanti problemi di raccordo sono risolti nella Staging Area attraverso l'impostazione di regole che definiscono:

- per ogni entità logica, indipendentemente dalla sua provenienza, l'insieme degli attributi e delle informazioni pertinenti che dovranno essere gestite;

- per ogni entità di dominio, le responsabilità permanenti dei dati, ovvero quale è il flusso alimentate che assume la funzione di master e al quale gli altri devono raccordarsi;
- per ogni anagrafica che presenti delle sovrapposizioni su più fonti, le responsabilità permanenti dei dati sui singoli attributi che creano conflitto
- eventuali responsabilità dei dati su base temporale (fa fede l'ultimo, il più recente, etc) o di qualità (fa fede la sorgente che porta i dati più completi, etc.)
- le regole di raccordo tra i flussi con funzione di master e quelli subalterni

Per soddisfare le esigenze di raccordo tra le fonti appena espresse, sono presenti all'interno dell'infrastruttura dati, una serie di tabelle cosiddette di Lookup mediante le quali si potrà costruire una sorta di associazione a più livelli tra le chiavi associate da parte di ogni sorgente al proprio record e la chiave stabilità in ambito di Anagrafica principale.

Strutture analoghe sono predisposte nella Staging Area per ogni dominio/anagrafica. La presenza ed il mantenimento di tali tabelle permette di velocizzare le procedure di ETL. Infatti, a fronte di ogni segnalazione, tali procedure, prima di applicare tutte le regole di verifica sopra descritte, verificano se il record da elaborare risulti già presente nel sistema accedendo alla tabella usando come chiave di ricerca la colonna relativa al proprio Sistema Sorgente.

Come linea guida complessiva, dei tre livelli di strutturazione dei dati (Staging Area, EDW, Data Marts) il luogo che soddisfa tutti i requisiti di consistenza, storicizzazione, normalizzazione e razionalizzazione dei dati è l'Enterprise Data Warehouse. La Staging Area gioca un ruolo ausiliare e preparatorio, i Data Marts un ruolo derivato e volto all'utilizzo. L'EDW mantiene e garantisce la completezza del contenuto informativo dove, un'ipotetica distruzione del livello di Data Marts potrebbe essere totalmente rigenerata a partire dall'EDW e la perdita dei flussi nell'Area di Staging, non ancora caricati in esso, potrebbe essere compensata da una ritrasmissione degli stessi da parte dei sistemi sorgenti.

3.4.7 Procedure ETL

L'ETL è un elemento infrastrutturale del data warehouse, che deve eseguire le trasformazioni richieste dalla complessità del business con adeguato livello di servizio e performance. Per descrivere le procedure di ETL, è utile fare riferimento allo schema di definizione delle fasi di elaborazione (si veda la Figura 3.3). In tale schema viene riportato il **legame logico delle varie fasi di elaborazione con l'architettura tecnica del data warehouse**. Le fasi d'elaborazione ETL si dividono in:

- **Acquisizione Dati**
- **Caricamento dell'Enterprise Data Warehouse**
- **Caricamento dei Data Marts**

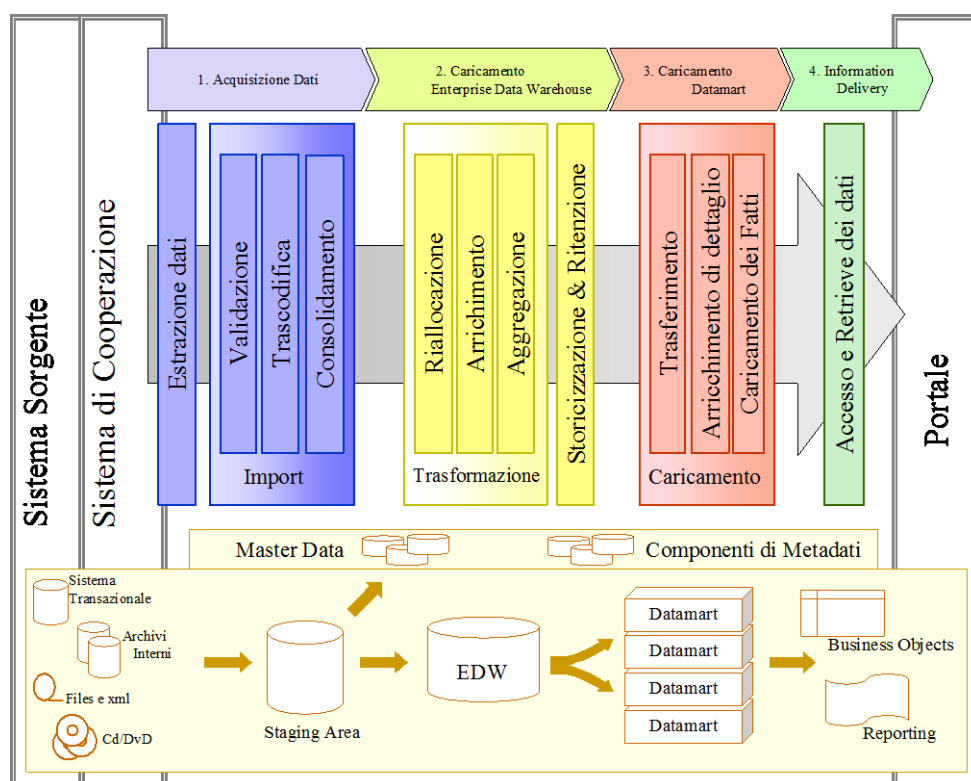
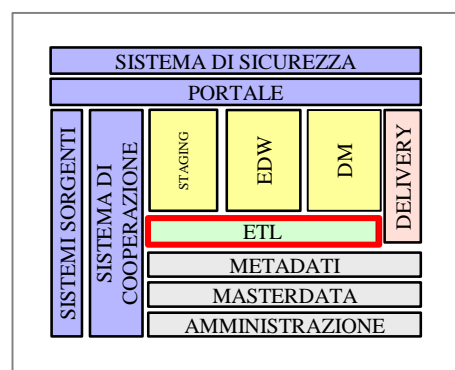


Figura 3.3 - Flussi d'Elaborazione

Nei capitoli che seguono si descrivono l'insieme delle singole funzionalità svolte dalle singole fasi previste. Ciò che va evidenziato in questo contesto è la necessità che l'intero processo di **Estrazione → Trasformazione → Caricamento** è *univocamente* definito ed implementato attraverso un *framework* comune a tutti i processi applicativi specifici dei singoli progetti.

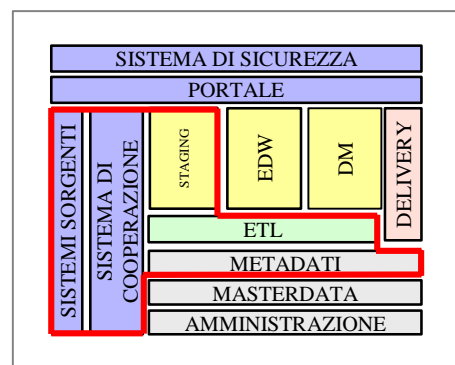
Esempi di servizi forniti da questo framework sono:

- gestione della fase di accoglienza e di elaborazione dei flussi;

- verifica esistenza in staging dei flussi da elaborare;
- verifica standard di nomenclatura e controllo validità formale flussi;
- generazione metadati tecnici relativi alla fase di acquisizione (logging);
- richiamo delle funzionalità applicative (processi ETL custom) per l'elaborazione dei flussi. Ciò comporta la necessità di definire in modo puntuale le interfacce e le modalità con cui i singoli processi di ETL potranno essere richiamanti;
- gestione della fase di logging;
 - fornitura di servizi (API) utilizzabili dai singoli processi di ETL per l'alimentazione dei metadati tecnici relativi allo stato dello specifico processo (logging);
 - tracciatura delle singole fasi elaborative (start, Stop, esito, ecc);
 - fornitura di funzionalità per la consultazione dei metadati tecnici.

3.4.7.1 Acquisizione Dati

Il flusso di “Acquisizione dei dati” prevede una serie di procedure costituite da un insieme di processi che permettono di valutare la consistenza delle informazioni e la gestione del trasferimento dei dati attraverso il sistema di cooperazione verso una base dati temporanea, la Staging Area. L’esecuzione di ogni procedura è definita attraverso un processo sequenziale che deve tenere conto delle dipendenze e dall’esito di ognuna. Di seguito vengono descritte in dettaglio (Figura 3.4):



- Procedura di Estrazione dei Dati.

Ha come effetto scatenante la generazione di un evento determinato da regole d’alimentazione che differiranno nella modalità e nella frequenza (eventi a pianificazione temporale o funzionale). La modalità (estrazione completa, estrazione per delta) e la frequenza dipendono dall’ente di provenienza del dato e dalla regola di ricezione applicata. L’estrazione consiste precipuamente in un trasferimento dei dati dal canale di cooperazione verso la staging area.

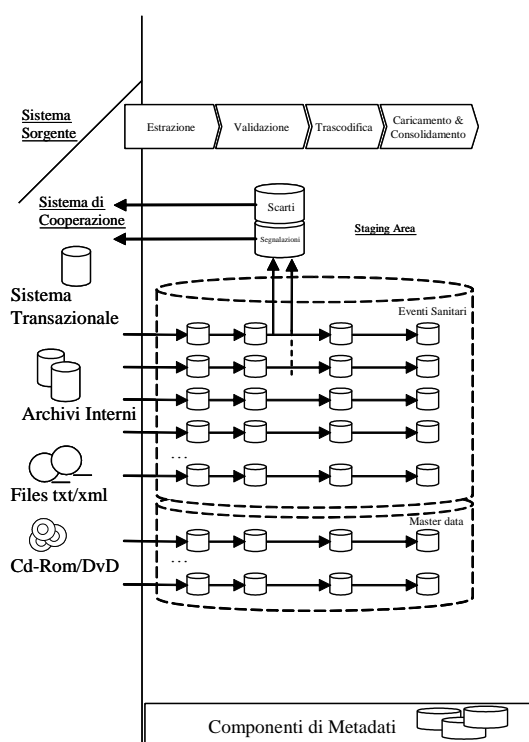


Figura 3.4 – Dettaglio Flusso di Acquisizione Dati

- Procedura di Validazione.

Vengono rilevate incoerenze ed inconsistenze dei dati forniti in input dalla procedura d'estrazione. La validazione avviene attraverso una serie di controlli:

1. Controllo dei dati a livello atomico
 - Controlli formali (controllo formale delle date, controllo formale dei campi numerici ed alfanumerici, controllo relativo all'obbligatorietà del dato, controllo formale del contenuto di alcuni dati)
 - Controllo dell'integrità referenziale (controllo coerenze delle informazioni interconnesse, controlli inerenti il trattamento di dati sensibili)
2. Controllo degli indicatori sullo stato dei dati
 - Verifica della completezza dei dati (attraverso regole d'alimentazione)
 - Verifica della tempestività di ricezione delle informazioni (attraverso il controllo dei tempi di ricezione)

Inoltre durante la procedura di validazione, in seguito alla generazione di errori e/o warnings, vengono inviati agli enti nei modi e nei tempi specificati, attraverso il sistema di cooperazione, le eventuali segnalazioni d'inconsistenza ed i dati scartati.

- Procedura di Trascodifica:

Attraverso questa procedura, sono implementate eventuali transcodifiche e decrittazione dei dati.

- Procedura di Consolidamento

Consolidamento delle informazioni nella Staging area.

A differenza dei sistemi operazionali, che ne consentono normalmente la gestione attraverso delle interfacce utente, l'aggiornamento delle strutture anagrafiche (Master data) nel Data Warehouse avviene attraverso la stessa fase di acquisizione dei dati, nei modi appena descritti.

3.4.7.2 Caricamento dell'Enterprise Data Warehouse

Il flusso di "Caricamento dell'EDW" prevede due procedure principali:

- Trasformazione
- Storizzazione/Ritenzione

Vediamole nel dettaglio (Figura 3.5):

- Procedura di Trasformazione:

Attraverso la trasformazione dei dati, **sono eseguite le riallocazioni, gli arricchimenti e le prime operazioni d'aggregazione**. Le modalità di trasformazione e del trattamento del dato dipendono dalle logiche descritte nelle regole dei metadata.

In particolare, nell'arricchimento i dati poveri (attributi di qualunque dimensione ritenuta opportuna all'analisi, come il tempo, la geografia,...) vengono "completati" in base alle informazioni presenti sulle tabelle anagrafiche.

Nelle procedure di riallocazione e aggregazione, viene solitamente presa in riferimento anche la dimensione temporale, attraverso il calendario di sistema (Principio della Competenza del dato, per cui i movimenti vengono consuntivati negli aggregati temporali di loro competenza). In questo modo è possibile tracciare separatamente gli eventi che, riferendosi ad un certo periodo, permettono di acquisire dati con differenti date di competenza al fine di completare periodi già consolidati o per la registrazione del periodo corrente.

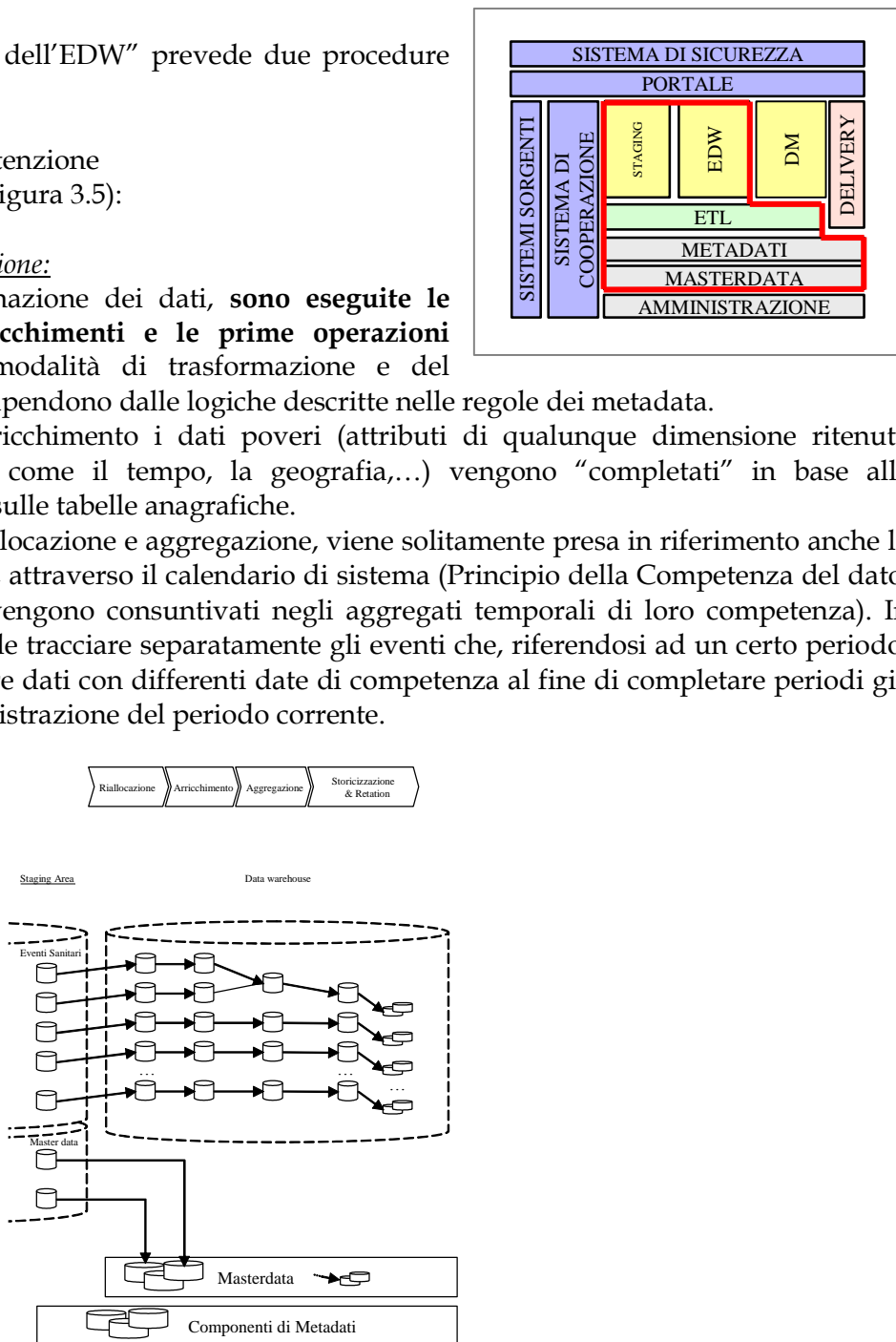


Figura 3.5 - Dettaglio Caricamento dell'EDW

- Procedura di Storizzazione e Ritenzione

Queste procedure permettono di gestire eventuali difficoltà in termini di moli di dati e di meccanismi di consuntivazione. In particolare la procedura di storizzazione permette di

introdurre un'informazione di "originalità" in riferimento al tempo. Vengono adottate due possibili soluzioni che faranno riferimento al principio di competenza delle informazioni:

- Gestione dello storico a valore assoluto, il movimento giunto in "ritardo" (cioè dopo la chiusura del periodo di sua competenza) viene consuntivato sia sul progressivo anno che sullo storico di sua stretta competenza.
- Gestione dello storico a chiusura, il movimento giunto in "ritardo" non viene consuntivato sullo storico e quindi viene caricato solo sul progressivo anno.

La procedura di ritenzione, invece, permette di eseguire la rimozione di dati appartenenti a periodi ritenuti "vecchi", per i quali non è più necessario conservare traccia d'alcun dato. Questa procedura è strettamente correlata alle procedure di backup dei dati.

La fase di trasformazione delle informazioni può impattare anche i master data.

- Normalizzazione delle anagrafiche; può essere di due nature:
 - Normalizzazione degli attributi di una stessa entità anagrafica: differenti attributi di una stessa entità possono essere separati in diverse tabelle.
 - Normalizzazione della relazione gerarchica tra differenti entità.
- Denormalizzazione delle anagrafiche; può prevedere due tipologie:
 - Denormalizzazione dei dati (merging).
 - Denormalizzazione della gerarchia (generazione dei figli variati).

L'esecuzione della fase di caricamento dell'EDW non dipende necessariamente dalla fase di "Acquisizione Dati" che la precede.

3.4.7.3 Caricamento dei Data Marts

Nel flusso di caricamento dei Data Mart i dati sono caricati in modo incrementale o totale secondo la strategia adottata. Nel caso di EDW distribuiti o di Data Mart il modulo di caricamento si occupa anche della distribuzione dei dati.

L'esecuzione di questa fase **comprende operazioni di aggregazione, denormalizzazione ed in alcuni casi viene eseguito un ulteriore arricchimento** (arricchimento di dettaglio) dei dati.

Il flusso di "Caricamento dei Data Mart" prevede una serie di procedure:

- Trasferimento dei dati
- Arricchimento di dettaglio
- Caricamento dei fatti

L'esecuzione di ogni procedura è definita attraverso un processo sequenziale che deve tenere conto delle dipendenze e dell'esito di ogni singola procedura. Queste procedure sono costituite da un insieme di processi (vedi Figura 3.6) che permettono di trasferire i dati opportunamente selezionati (es: per periodo di competenza), attraverso una semplice copia massiva in una prima area del Data Mart definito come "Livello Base". Il trasferimento dei dati così effettuato permette di minimizzare i tempi di accesso all'enterprise data warehouse e di rendere più performanti l'insieme dei processi che vi accedono in parallelo. Successivamente attraverso procedure di arricchimento e caricamento dei dati vengono determinate tutte le informazioni di dettaglio in una seconda area definita come "Livello Derivato". Il livello derivato del Data Mart costituisce l'area preposta all'information delivery.

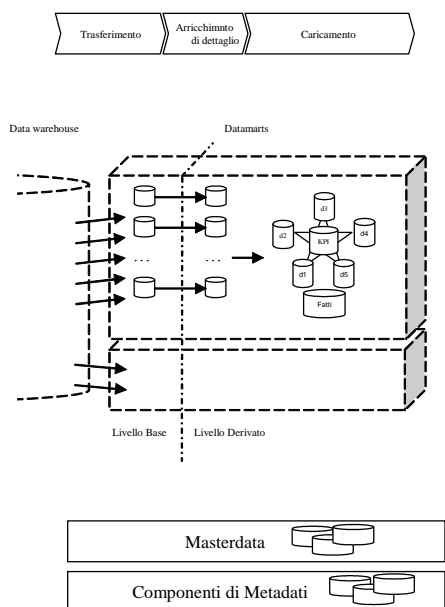


Figura 3.6 - Caricamento dei Data Mart

Al fine di rendere più agevoli i processi di ricerca ed estrazione dei dati e, di ridurre la numerosità delle righe, le operazioni di aggregazione e denormalizzazione degli attributi differenti di una stessa entità, possono essere riuniti nella stessa tabella sulla base di criteri di numerosità ed

omogeneità e si possono spostare sulle colonne cause differenti di una stessa tipologia di movimento.

Per la denormalizzazione e l'aggregazione dei dati è possibile utilizzare due metodi:

- *metodo a "pettine"*: l'aggiornamento avviene in parallelo sulle strutture di Bottom Level e sulle strutture aggregate. Questo tipo di metodo permette:
 - Carico limitato e uniforme sui rollback segment
 - Numero di transazioni al DB limitato al minimo (minima movimentazione di dati e massimo uso della memoria, performance lineari con la mole di dati processata)
 - Agevole gestione del tempo
 - Flessibile validazione dei dati (possibilità di gestire scarti in modo oculato)
 - Flessibilità nella gestione di tabelle aggregate con strutture diverse da quelle di dettaglio
 - Possibilità di forte parallelizzazione senza mettere in crisi il sistema.

Gli svantaggi:

- Sort necessari per avere performances accettabili su alcune aggregazioni (richiesta di memoria e di spazi)
- Performances inferiori in certe condizioni (soprattutto su tabelle molto aggregate) dove il sort è determinante

- *metodo a "stella"*: l'aggiornamento avviene prima sulle strutture di Bottom Level (dove vengono memorizzati i dati di dettaglio) e successivamente a partire da queste, sulle strutture aggregate. Questo tipo di metodo permette:

- Maggiore semplicità di implementazione
- Assenza di sort e quindi minore utilizzo di memoria e disco.
- Performances specifiche superiori nel caricamento delle tabelle aggregate

Gli svantaggi:

- Maggiore uso di transazioni su DB: degrado performance + che lineare
- Maggiore rigidità e quindi minori possibilità di "ragionare" sui dati in fase di caricamento
- Carico di punta su rollback e conseguente difficoltà nel parallelizzare operazioni pesanti

Consolidamento dei dati

La procedura di consolidamento permette d'integrare ed omogeneizzare varie classi di dati, da cui attingere per produrre output confrontabili e consistenti.

I metodi per la definizione delle procedure di consolidamento fanno solitamente riferimento agli attributi del periodo, considerando quindi operazioni di:

- Aggregazione dei dati in funzione del tempo
- Definizione degli indicatori a livello cumulato (year-to-date, year-to-month)
- Ricostruzione delle informazioni storiche (riallocazione/riclassificazione)

La gestione del caricamento di dati a livello aggregato è un problema comune a tutti i Data Mart, infatti alcune informazioni possono nascere e quindi essere disponibili solo a livello aggregato. In particolare possiamo distinguere due tipi d'informazioni aggregate che dovranno essere gestite:

1. per le informazioni che arrivano solo a livello aggregato i valori si devono consolidare.

2. per le informazioni che arrivano sia a livello aggregato che a livello elementare i valori non si devono consolidare

Il consolidamento delle informazioni è più semplice e lineare se parte sempre dalle tabelle di massimo dettaglio; a questo scopo sono trattate su tali tabelle anche le informazioni che pervengono aggregate tramite l'introduzione di elementi "dummy". Ogni movimento che arriva non a livello elementare, è direttamente associato al suo articolo "dummy".

Nella definizione delle procedure di consolidamento, si è tenuto conto dei requisiti dell'analisi funzionale per identificare metodi e tempi da considerare nelle differenti aree funzionali del dato. Il consolidamento, è definito sulle strutture di dettaglio (nell'enterprise data warehouse) attraverso procedure di "append", e sulle strutture aggregate (nei Data Marts) attraverso procedure di "update".

3.4.8 Integrazione

Il sottosistema di DW sin qui descritto deve, in generale, rispondere agli stessi requisiti di integrazione richiesti per gli altri sottosistemi dell'infrastruttura tecnologica e, in particolare, con la componente di sicurezza.

Tali aspetti riguardano in prima istanza l'integrazione degli strumenti di delivery costituiti da prodotti di terze parti con l'infrastruttura di sicurezza e profilazione stabiliti per SIS-N.

Nello specifico del contesto del SIS-N, le problematiche di integrazione degli strumenti di Query & Reporting e più in generale delle funzionalità di OLAP e Data Mining possono essere ricondotte alla capacità sostanziale di poter utilizzare tali strumenti in modo da raggiungere un livello di integrazione che risulti coerente con le linee guida e le scelte architetturali già attuate in ambito di sicurezza.

Il problema della sicurezza non si pone a livello architetturale nel sistema di Data Warehouse in quanto è appunto a carico dell'infrastruttura di sicurezza che sottende a tutto il progetto SIS-N il garantirne un perimetro certificato e controllato. L'aspetto della sicurezza che qui interessa maggiormente riguarda invece il grado e livello di accessibilità ai dati ed ha pertanto una forte relazione con gli aspetti più propriamente inerenti la profilatura delle utenze.

Il primo requisito da garantire è quello del Single Sign On nella relazione con tutte le applicazioni del SIS-N. A questo scopo, lo strumento *Authorization Profile Manager* costituisce l'elemento con funzione di garante rispetto alle credenziali dell'utente e alle sue abilitazioni funzionali.

In questa profilatura di primo livello sono bilanciate le abilitazioni a livello di macro-utilizzo delle funzionalità offerte dagli applicativi realizzati sul Data Warehousing ma non è possibile pensare qui ad una profilatura di dettaglio sui reports o sulla visibilità specifica all'interno di ciascuno di essi. Questi sono infatti aspetti imprescindibilmente legati al packaged utilizzato ed implicano un livello informativo troppo dettagliato e contestualizzato per inserirsi a livello dell' *Authorization Profile Manager*. Vi è invece una divisione delle responsabilità dove:

- *Authorization Profile Manager*: ha la responsabilità di gestione delle credenziali dell'utente, l'abilitazione alle macro-funzionalità di utilizzo ed interrogazione del Data Warehouse. Fornisce anche tutte le informazioni proprie dell'utente necessarie per limitare la visibilità sui dati nel Data Warehouse (ex. ruolo, unità organizzativa, etc.).
- *Sistema di profilatura della piattaforma per la reportistica*: gestisce la profilatura di dettaglio, attribuendo a ciascun utente/ruolo regole di gestione e visibilità sui singoli report e le opzioni di filtro sui dati in essi mostrati.

L'accesso agli applicativi di interrogazione e di reportistica sarà trasparente per l'utente ma non esaurisce la necessità di dover censire sul repository di prodotto le singole utenze/ruoli.

E' quindi necessario:

- propagare la struttura di utenze/ruoli dal repository dell' *Authorization Profile Manager* a quello specifico di prodotto. Chiaramente non si intende una duplicazione della profilatura ma solo la duplicazione delle utenze, senza credenziali e senza attributi di dettaglio
- definire quali attributi dell'utente, eventualmente in aggiunta al ruolo, incidano sulla visibilità dei dati

A questo punto il controllo si sposta sul piano proprio del Data Warehouse, dove per regolare la visibilità sui dati è necessario (oltre all'avere un repository fornito e gestito dal prodotto specifico):

- censire la reportistica puntualmente o secondo regole
- regolare la visibilità sulla reportistica in generale
- regolare la visibilità sui dati mostrati

4 ARCHITETTURA TECNICA NEGLI SCENARI EVOLUTIVI IDENTIFICATI

La capacità di evolvere è essenziale per un'infrastruttura di Data Warehouse e questo perché il sistema deve essere pronto a rispondere alle nuove esigenze/aspettative degli utenti.

Ovviamente una soluzione che riesca a conciliare tali esigenze dovrà necessariamente porre una serie di vincoli e/o compromessi che in una qualche natura condizioneranno la futura evoluzione del sistema.

4.1 Architettura Tecnica

La soluzione tecnica implementata per il sottosistema nazionale DW prevede in sintesi l'utilizzo delle seguenti componenti software:

- RDBMS
- Sistema Operativo
- Piattaforma di Info Delivery
- Piattaforma di ETL

E' schematizzata nel seguito la collocazione dei singoli prodotti sulle diverse componenti funzionali individuate per un'infrastruttura software

- RDBMS: **Oracle12c**;
- Sistema Operativo: **Unix**;
- Piattaforma di Info Delivery: application server **Business Object 4.2** /business analytics **IBM SPSS**;
- Piattaforma di ETL: infrastruttura **PL/SQL** e **Java**, **SAP DataServices 4.2**

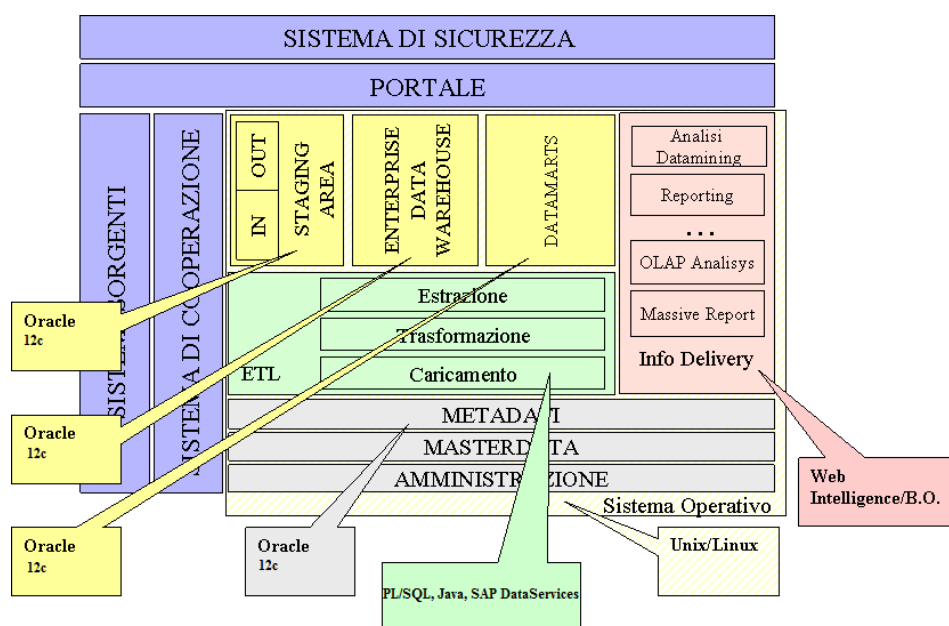


Figura 4.1 - Architettura Tecnica

4.2 Architettura Operativa

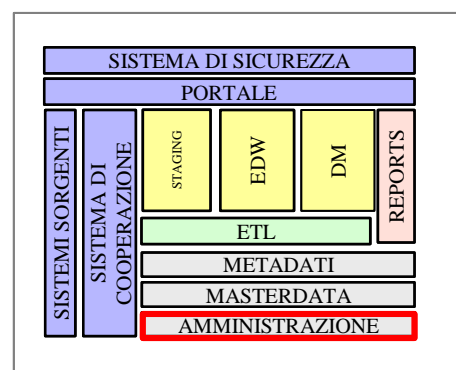
L'architettura Operativa, rappresentata dal modulo di amministrazione, è costituita da un insieme di procedure che permettono il supporto e la manutenzione applicativa dell'architettura data warehouse.

4.2.1 Procedure di Ritenzione dei dati

Ogni volta che viene effettuata una "chiusura" di un periodo, si avrà contemporaneamente la rimozione di dati appartenenti a periodi ritenuti "vecchi", per i quali cioè non è più necessario (a livello di controllo) conservare traccia d'alcun dato.

Il numero di periodi che voglio mantenere in "vita" è strettamente legato al tipo di controllo che deve essere effettuato. Se per esempio è utile avere un confronto dei dati su anni passati, è evidente che il numero di periodi che si devono mantenere attivi deve essere uguale almeno a dodici mesi.

Il numero di periodi in linea su ciascuna tabella è mantenuto ad un valore soglia da procedure di pulizia periodiche. La permanenza dei dati storici sulla struttura di memorizzazione deve bilanciare il trade off tra esigenza di reperimento dei dati e gestibilità degli stessi in seguito all'aumentare della mole.



4.3 Architettura del Modello dati

Come descritto nelle sezioni precedenti, la struttura del modello dati del sistema di DW del Nodo nazionale di SIS-N è costituito da:

- un'area preposta alla ricezione ed elaborazione dei dati (*Staging Area*);
- un'area in cui sono memorizzati l'insieme dei metadati necessari al monitoraggio tecnico ed alla comprensione dei dati gestiti nell'ambito del sistema (*Metadati*);
- un'area contenente i riferimenti Anagrafici relativi a tutte le entità concettuali presenti nel sistema (*MasterData*) utilizzata per eseguire le funzione di validazione e di normalizzazione dei dati acquisiti;
- un Enterprise Data Warehouse;
- una serie di DataMart tematici;

Come si può facilmente intuire, gli strati del modello dati che risultano particolarmente critici sono costituiti rispettivamente da:

- lo strato di Enterprise Data Warehouse;
- la componente di MasterData;

Tali livelli, infatti, dovrebbero essere strutturati in modo da poter soddisfare le esigenze informative delle diverse esigenze di Business che nel tempo saranno implementati nell'ambito del sistema. In termini generali, esistono due diverse macro tipologie di approcci per la costituzione di un sistema di DataWarehouse:

- **approccio top-down:** in cui si prevede la costruzione di un modello dati di tipo Enterprise. In questo scenario, lo sviluppo dei DataMart è condizionato alla presenza preventiva di uno strato consolidato di dati normalizzati all'interno dell'EDW. Questa modalità presenta quale aspetto negativo lo svantaggio che i singoli progetti implementativi devono essere preceduti da un passo di modellizzazione iniziale atto a definire il modello dati dell'intero sistema SIS-N;
- **approccio incrementale:** in cui il modello dati complessivo viene costruito a seguito di cicli di consolidamento eseguito a valle della definizione di singoli DataMart. Il punto di criticità di questo approccio è rappresentato dalla capacità di poter opportunamente governare il processo di riciclo/consolidamento.

Tra le due filosofie implementative descritte, nell'ambito del contesto del SIS-N è stata adottata la strategia di sviluppo incrementale coniugata nell'ambito della metodologia di implementazione denominata La "BUS ARCHITECTURE" di R. Kimball (vedi Appendice B per i dettagli), che si basa sulla costituzione del sistema di DW analizzando le diverse esigenze di business. Il metodo proposto da Kimball porta difatto alla costituzione di una serie di "Datawarehouse Tematici" comprendenti sia la componente di dettaglio sia le relative strutture aggregate. Tale infrastruttura costituisce una sorta di "BUS di dati" (la "BUS ARCHITECTURE") che arricchendosi di strutture potrà essere utilizzata anche dalle altre applicazioni.

Tuttavia, indipendentemente dall'approccio individuato per la costruzione dell'EDW, particolare attenzione è stata posta nella fase di modellizzazione del livello di MasterData che contiene l'insieme di tutte le strutture Anagrafiche utilizzabili per valicare e normalizzare i dati acquisiti ed

inseriti nell'EDW. Per tale livello, si è reso necessario utilizzare un approccio che pur non essendo totalmente di tipo top-down doveva necessariamente essere concepito in modo tale da assicurare il massimo livello di flessibilità al fine di minimizzare e ridurre i ricicli a seguito di nuove esigenze che dovessero emergere a valle dell'analisi di obiettivi di business relativi a nuovi progetti implementativi.

5 APPENDICI

APPENDICE A - Flusso Logico dei Dati

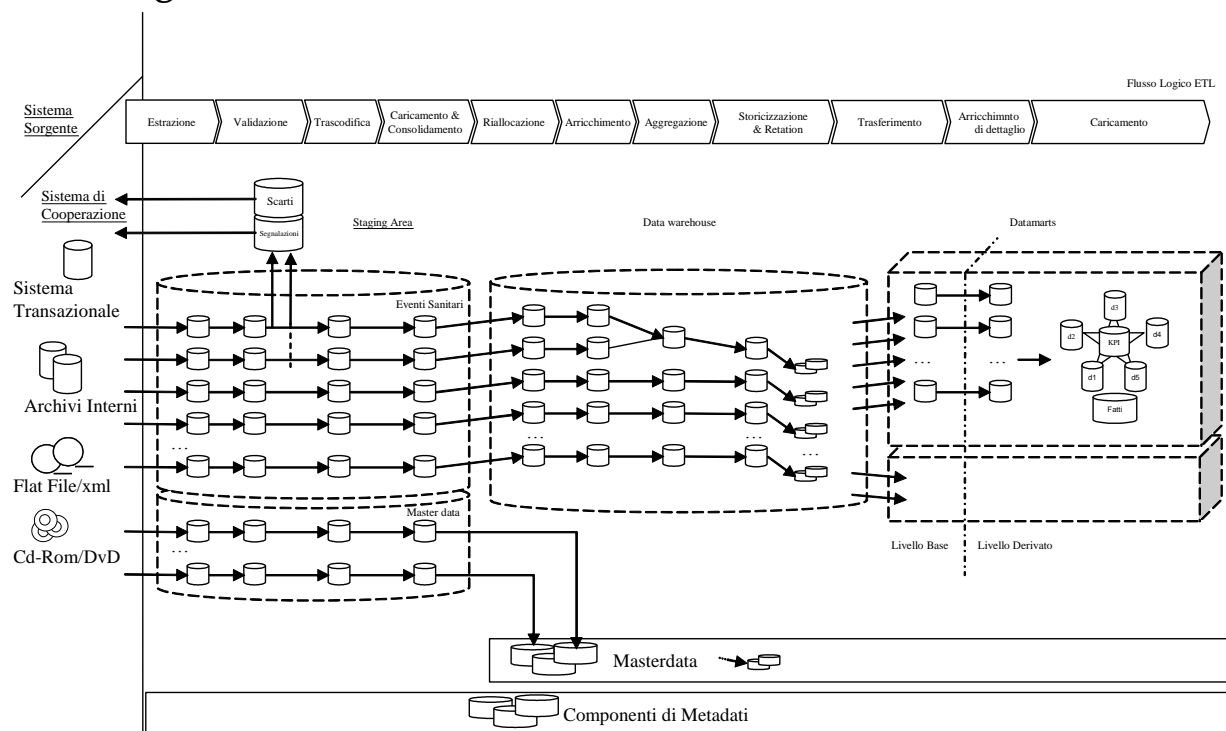


Figura 5.1 - Processo Logico ETL

APPENDICE B – Data Warehouse, Modelli Concettuali di Riferimento

In letteratura esistono diversi modelli concettuali di data warehousing. Da tali modelli astratti derivano i modelli reali che trovano applicabilità nei progetti implementativi.

Tra questi si può fare riferimento agli studi di W.H.Inmon e R. Kimball, i cui modelli costituiscono ad oggi i due approcci dominanti all'enterprise warehousing.

La "CORPORATE INFORMATION FACTORY (CIF)" di W.H. Inmon

L'architettura di W.H. Inmon mostrata in Figura 5.2 è denominata Corporate Information Factory (nel seguito del capitolo ci riferiremo ad essa con l'abbreviazione CIF).

I dati vengono inizialmente estratti dai sistemi sorgente per essere passati ad un'area di staging. Questa area viene caricata con dati atomici.

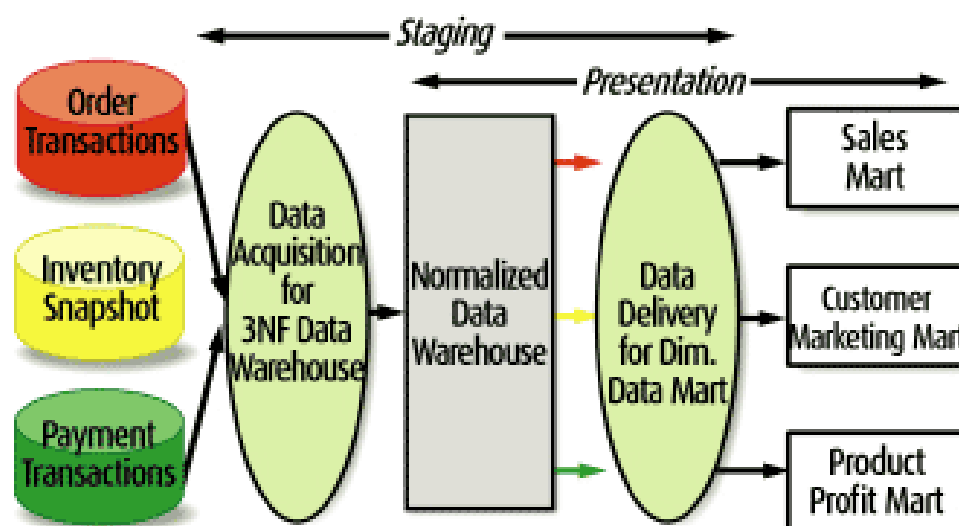


Figura 5.2 – Corporate Information Factory

Tali dati subiscono un processo di normalizzazione e sono inseriti nel Data Warehouse per essere poi passati ai datamarts dimensionali.

In questo scenario i datamarts sono costruiti a partire dai dipartimenti di business, con dati aggregati e strutturati dimensionalmente. I dati atomici sono accessibili attraverso il data warehouse normalizzato. Ovviamente la struttura dei dati atomici sarà notevolmente differente da quella delle informazioni riassunte nei datamarts.

La "BUS ARCHITECTURE" di R. Kimball

L'architettura di Kimball è denominata "BUS ARCHITECTURE". Vediamo nel dettaglio la teoria che lo caratterizza: i dati provenienti da sorgenti esterne sono trasformati in informazioni presentabili nella staging area. Le operazioni di staging iniziano con l'estrazione dai sistemi operazionali sorgenti, successivamente i dati vengono trasformati in modo opportuno per poter essere passati all'area di presentazione. Alcune attività svolte nell'area di staging sono centralizzate come ad esempio lo storage di referenze comuni dei dati, mentre altre possono essere distribuite.

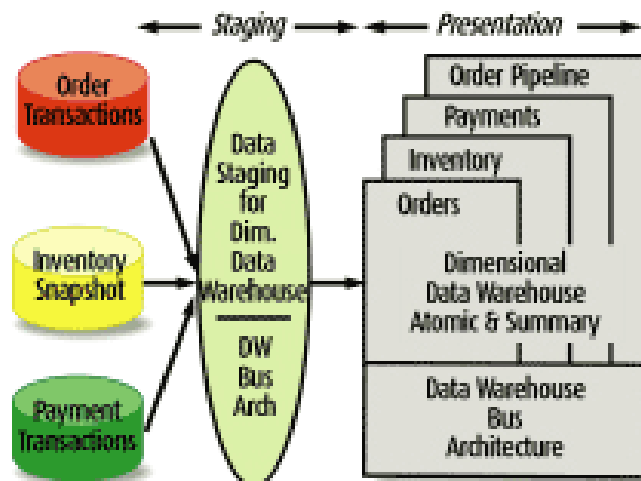


Figura 5.3 - BUS Architecture

Come si evince dalla Figura 5.3 l'area di presentazione è dimensionalmente strutturata. Un modello dimensionale contiene le stesse informazioni di un modello normalizzato ma i diversi pacchetti sono studiati per essere di facile utilizzo ed offrire performance nelle query.

L'area di presentazione include sia dettagli atomici che informazioni riassunte (aggregate in tabelle relazionali o cubi multidimensionali) al fine di ottenere buone performance ed una distribuzione geografica dei dati. Le queries discendono progressivamente a livelli sempre più bassi di dettaglio, senza il bisogno di parametrizzazioni dedicate da parte dell'utente.

In questo approccio i modelli dimensionali sono costruiti a partire dai processi di business (corrispondono a misurazioni o eventi di business), e non dai diversi dipartimenti. Ad esempio i dati riguardanti gli ordini sono prima caricati nel data warehouse dimensionale per un accesso di tipo enterprise, solo in seguito vengono replicati nei datamarts dipartimentali. I dati consolidati definiscono la matrice dei processi di business. Il BUS dell'enterprise data warehouse identifica e rafforza le relazioni tra le metriche dei processi di business (fatti) e gli attributi descrittivi (dimensioni).

DIFFERENZE SPECIFICHE DEI DUE MODELLI

Ci sono due fondamentali differenze tra la "Bus Architecture" di Kimball ed il "CIF" di Inmon. La prima riguarda la presenza nel CIF di una struttura dati normalizzata prima del caricamento dei modelli dimensionali. La seconda è nel trattamento dei dati atomici, nel CIF i dati atomici devono essere memorizzati nel data warehouse normalizzato, nell'approccio di Kimball invece i dati atomici vengono dimensionalmente strutturati.

APPROCCIO IBRIDO "Inmon" "Kimball"

Un terzo modello consiste in un approccio ibrido tra le due architetture appena descritte. Teoricamente un approccio di questo tipo dovrebbe unire il meglio del CIF e del BUS Architecture. Come si vede dalla Figura 5.4 è presente il data warehouse normalizzato del CIF (Inmon), più un data warehouse dimensionale contenente sia i dati atomici che quelli riassunti (basato quindi sul modello di Kimball).

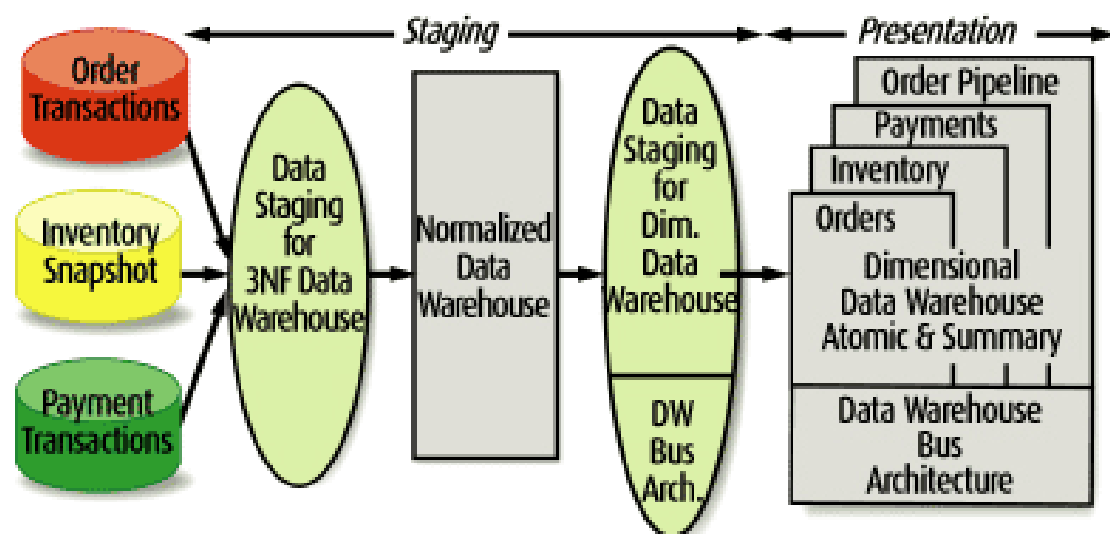


Figura 5.4 – Architettura Ibrida

APPENDICE C – Metadati per aree del Data Warehouse

Per ogni metadato sono fornite descrizione circa:

- Il nome del metadato;
- la natura: se fornito in input (IN) nel flusso, se generato dalla procedura di ETL (OUT);
- la fonte responsabile della valorizzazione del dato;
- una breve descrizione che riepiloga la funzione svolta dal metadato nell'ambito del processo di DW

1. Staging Area

I metadati riferiti ai dati ospitati sulla Staging Area e al processo ETL che si occupa di alimentarla a partire dai dati flussi presenti sulla porta di dominio hanno in particolare il compito di gestire il controllo e la validazione dei flussi (nella loro totalità più che per singola riga o attributo) e di risolvere i primi problemi di sincronizzazione e trasformazione. Nell'architettura prevista per il sottosistema nazionale DW, la Staging Area viene utilizzata non solo per acquisire i dati provenienti dai sistemi alimentanti, ma anche per ospitare eventuali flussi che, generati dalle procedure operanti all'interno del sistema di DW, devono essere spediti (attraverso l'infrastruttura di cooperazione del SIS-N) ai diversi utenti. Tali flussi possono essere suddivisi in:

- flussi contenenti segnalazioni di errore necessari per informare i Sistemi alimentanti Regionali sull'esito dell'elaborazione dei dati da essi forniti;
- flussi contenenti informazioni generate dal sistema di DW che devono essere fornite ad eventuali Amministrazioni che necessitano di ricevere informazioni statistiche sul patrimonio informativo del SIS-N (ISTAT, INAIL, Regioni,).

Metadati tecnici

Metadato	IN/OUT/ ELAB	Responsabilità del dato	Note
Nome del Flusso	IN	Sistema sorgente	Identificativo del flusso generato dal sistema sorgente ed acquisito tramite l'infrastruttura di cooperazione
Tipologia Flusso	IN	Sistema sorgente	Identificativo della tipologia di informazioni presenti nel flusso
Numero Record presenti nel Flusso	IN	Sistema sorgente	Numero degli oggetti presenti all'interno del flusso
Destinatario del Flusso	OUT	SIS-N	Identificativo del destinatario a cui deve essere spedito il flusso generato
Data aggiornamento	IN	Sistema sorgente	Data dell'ultimo aggiornamento dei dati contenuti nel flusso, perché non necessariamente i dati vengono estratti nel momento stesso in cui variano
Data estrazione	IN	Sistema sorgente	Timestamp dell'estrazione dell'intero flusso
Data ultima elaborazione	OUT/EL AB	ETL vs Staging	Solo nel caso in cui i dati elaborati che passano dalla staging area al dwh non vengano cancellati (Ex. per calcolo delta)
Dati di auditing	OUT	ETL vs Staging	Sono misure di prestazione del processo ETL eseguito (tempo di

Metadato	IN/OUT/ ELAB	Responsabilità del dato	Note
prestazionale			esecuzione, tempo di I/O, ecc.)
Dati di auditing qualitativo	OUT	ETL vs Staging	Sono misure sulla qualità dei dati rilevata nell'esecuzione del processo ETL (numero di record scartati, tipologie di errori riscontrati, ecc.)
Dati per il controllo e la gestione della qualità	IN	ETL vs Staging	Dati ausiliari per il controllo. Da definire quando saranno chiari i controlli sulla qualità. È la parte di definizione dei controlli.
Esiti dei controllo di flusso	OUT	ETL vs Staging	Flusso valido, scartato, scarti parziali e motivi degli scarti; da dettagliare sul singolo controllo quando saranno delineati quelli da effettuare
Fase prossima elaborazione	ELAB	ETL vs Staging	Prossima fase nel ciclo dei controlli/trasformazioni; da dettagliare sul singolo controllo quando saranno delineati quelli da effettuare
Fasi di elaborazione	IN	ETL vs Staging	Flow dei controlli e trasformazioni a cui sottoporre il flusso; da dettagliare sul singolo controllo quando saranno delineati quelli da effettuare
Formato del flusso	IN	Sistema sorgente / Sistema di cooperazione / ETL vs Staging	Formati e dimensioni
Logs	OUT	ETL vs Staging	Log Applicativo generato dalla procedura di ETL in cui vengono memorizzati i singoli eventi ritenuti rilevanti ai fini delle successive fasi di tuning e controllo
Regole di Mappatura	ELAB- OUT	ETL vs Staging	Mappatura tra sorgente-staging
Motivo degli scarti	OUT	ETL vs Staging	Da dettagliare sul singolo controllo quando saranno delineati quelli da effettuare
Periodicità di aggiornamento	IN	Sistema sorgente	Periodicità, orario, data di riferimento
Procedura di trasporto in ingresso	IN	Sistema sorgente / Sistema di cooperazione / ETL vs Staging	Nome della procedura
Procedura di trasporto in uscita	IN	Sistema di cooperazione / ETL vs Staging	Nome della procedura
Regole di accorpamento	IN	ETL vs Staging	Regole per l'accorpamento dei flussi dove questi provengano frazionati o da più sorgenti. Riguardano sempre un trattamento tecnico dei flussi
Regole di pulizia e normalizzazione	IN	ETL vs Staging	Sono regole per il trattamento tecnico dei flussi
Regole di trasformazione ed arricchimento	IN	ETL vs Staging	Sono regole per il trattamento tecnico dei flussi
schema logico e fisico se si tratta di table db	IN	dbms	Nomi, formati, relazioni. Implicito nelle tabelle di sistema

Metadato	IN/OUT/ ELAB	Responsabilità del dato	Note
Sincronizzazione	IN/ELAB	ETL vs Staging	Flusso che deve attendere, flusso che va a completare, ovvero tutte le informazioni che servono a sincronizzare il flusso in ingresso rispetto ad altri flussi in ingresso, al fine di poterli elaborare correttamente nelle fasi successive
Soglie di tolleranza	IN	ETL vs Staging	Soglie per gli specifici controlli, sui flussi interi
Sorgente del dato	IN	Sistema sorgente / Sistema di cooperazione	Nome del flusso, responsabilità; verranno dettagliati quali attributi
Stato del flusso	ELAB/OUT	ETL vs Staging	In arrivo, da validare, validato, da elaborare, ecc.. Da dettagliare sul singolo controllo quando sarà chiaro il processo di verifica e trasformazione
Tempi di elaborazione	OUT	ETL vs Staging	Durata delle singole fasi dell'elaborazione; da dettagliare sul singolo controllo quando saranno delineati quelli da effettuare
Tipologia di flusso	IN	Sistema sorgente	Flusso elementare/aggregato/primitivo/derivato

Metadati di business

Trattandosi di una fase di gestione puramente tecnica del dato, non si rilevano attributi di business rilevanti per la Staging Area.

2. Enterprise Data Warehouse

I metadati dei dati ospitati nell'Enterprise Data Warehouse, e dei processi ETL che lo alimentano, si riferiscono più in dettaglio ad ogni singola istanza presente sui vari flussi/tabelle in ingresso, dettagliando a livello di attributo le informazioni necessarie.

Configurano e mantengono tutte le informazioni relative alle regole di controllo, arricchimento, trasformazione, aggregazione e derivazione dei singoli attributi/entità.

Metadati tecnici

Metadato	IN/OUT/ELAB	Responsabilità del dato	Note
Nome del Flusso	IN	Sistema sorgente	Identificativo del flusso generato dal sistema sorgente ed acquisito tramite l'infrastruttura di cooperazione
Tipologia Flusso	IN	Sistema sorgente	Identificativo della tipologia di informazioni presenti nel flusso
Numero Record presenti nel Flusso	IN	Sistema sorgente	Numero degli oggetti presenti all'interno del flusso
Data aggiornamento	OUT/ELAB	ETL di alimentazione EDW	Data di ultima elaborazione corretta, che ha portato al caricamento del dato sull'EDW
Data elaborazione	OUT/ELAB	ETL di alimentazione EDW	Data dell'ultima elaborazione, anche se non terminata
Data presa in carico	OUT/ELAB	ETL di alimentazione EDW	Data in cui L'ETL di alimentazione del EDW ha preso in carico il dato per portarlo nell'EDW. Serve nel caso l'azione non sia stata portata a termine ma abbia raggiunto risultati parziali.
Dati di auditing prestazionale	OUT	ETL di alimentazione EDW	Sono misure di prestazione del processo ETL eseguito
Dati di auditing qualitativo	OUT	ETL di alimentazione EDW	Sono misure sulla qualità dei dati rilevata nell'esecuzione del processo ETL
Dati per il controllo e la gestione della qualità	IN	ETL di alimentazione EDW	Dati ausiliari per il controllo. Da definire quando saranno chiari i controlli sulla qualità. È la parte di definizione dei controlli.
Dati per il controllo e la gestione della qualità	IN	ETL di alimentazione EDW	Dati ausiliari per il controllo. Da definire quando saranno chiari i controlli sulla qualità. È la parte di definizione dei controlli.
Destinatario del dato	IN/ELAB	ETL di alimentazione EDW	Table dell'EDW target
Esiti dei controllo	OUT/ELAB	ETL di alimentazione EDW	Dato valido, scartato, scarti parziali e motivi degli scarti; da dettagliare sul singolo controllo quando saranno delineati quelli da effettuare
Fase prossima elaborazione	ELAB	ETL di alimentazione EDW	Prossima fase nel ciclo dei controlli/trasformazioni. Da dettagliare sul singolo controllo quando saranno delineati quelli da effettuare
Fasi di elaborazione	IN	ETL di alimentazione EDW	Flow dei controlli e trasformazioni. Da dettagliare sul singolo controllo quando saranno delineati quelli da effettuare
Formato dei dati	IN	modellazione dati	Formati e dimensioni
Logs	OUT	ETL di alimentazione	Log Applicativo generato dalla procedura di ETL in cui

Metadato	IN/OUT/ ELAB	Responsabilità del dato	Note
		EDW	vengono memorizzati I singoli eventi ritenuti rilevanti ai fini delle successive fasi di tuning e controllo
Regole di Mappatura	ELAB- OUT	ETL di alimentazione EDW	Mappatura tra staging-EDW. Per singolo campo l'associazione con il flusso nella staging o, se questa è soggetta a cancellazione, con la sorgente esterna
Motivo degli scarti	OUT/ELAB	ETL di alimentazione EDW	Dettagliato per ogni errore; da dettagliare sul singolo controllo quando saranno delineati quelli da effettuare
Periodicità di aggiornamento	IN	ETL di alimentazione EDW	Periodicità, orario, data di riferimento
Procedura di trasporto in ingresso	IN	ETL di alimentazione EDW	Tra Staging e EDW
Procedura di trasporto in uscita	IN	ETL di alimentazione EDW	Per gli scarti verso la Staging
Regole di aggregazione e di calcolo	IN	ETL di alimentazione EDW	Sono regole per il trattamento tecnico dei dati
Regole di trasformazione ed arricchimento	IN	ETL di alimentazione EDW	Sono regole per il trattamento tecnico dei dati
schema logico e fisico se si tratta di table db	IN	dbms	Nomi, formati, relazioni. Implicito nelle tabelle di sistema
Sincronizzazione	IN/ELAB	ETL di alimentazione EDW	Parametri e semafori per la sincronizzazione delle varie procedure di trasformazione e trattamento dei dati
Soglie di tolleranza	IN	ETL di alimentazione EDW	Soglie per gli specifici controlli sui dati puntuali
Stato del dato	ELAB	ETL di alimentazione EDW	In arrivo, da validare, validare, da elaborare, ecc.. Da dettagliare sul singolo controllo quando sarà chiaro il processo di verifica e trasformazione
Tempi di elaborazione	OUT	ETL di alimentazione EDW	Durata delle singole fasi dell'elaborazione. Da dettagliare sul singolo controllo quando saranno delineati quelli da effettuare
Tipo dato/attributo	IN	ETL di alimentazione EDW	Dato elementare/aggregato/primitivo/derivato

Metadati di business

Metadato	IN/OUT/ ELAB	Responsabilità del dato	Note
Aggregazione a cui partecipa	IN	ETL di alimentazione EDW	(vuote)
dato presente/aggiornato	ELAB/OUT	ETL di alimentazione EDW	Per tutti i dati di business di cui si vuole evidenziare la presenza. Ex. esiste il saldo contabile? É caricato ? É aggiornato e corretto?
definizione delle misure	IN	ETL di alimentazione EDW	(vuote)
Formule di calcolo e derivazione	IN	ETL di alimentazione EDW	Sono regole per il trattamento dei dati per finalità di business.
KPI	IN	ETL di alimentazione EDW	(vuote)
Periodo di validità	ELAB/OUT	ETL di alimentazione EDW	(vuote)
Profili utenti e regole di visibilità	IN	ETL di alimentazione EDW	(vuote)
Regole di business	IN	ETL di alimentazione EDW	Sono regole per il trattamento dei dati per finalità di business.
Regole di confronto e trattamento dei dati legati alla loro valore semantico e al significato di business	IN	ETL di alimentazione EDW	(vuote)
regole di validazione	IN	ETL di alimentazione EDW	Controlli secondo logiche di business
Ritardo	ELAB/OUT	ETL di alimentazione EDW	con quale ritardo è entrato nel sistema
Storicizzato ?	IN	ETL di alimentazione EDW	(vuote)

3. Data Marts

L'Enterprise Data Warehouse costituisce il livello di normalizzazione e strutturazione dei dati e la sua alimentazione è soggetta a diverse elaborazioni che possono produrre scarti. I data mart si innestano su di esso e quindi in una situazione già ordinata, normalizzata e con dati validi per cui gli scarti e le situazioni di errore dovrebbero essere a regime nulle.

Si presentano in maniera preponderante i metadati di business, proprio perché è la fase rivolta esplicitamente all'utenza finale.

Metadati tecnici

Metadato	IN/OUT/ELAB	Responsabilità del dato	Note
Data caricamento	OUT	ETL di alimentazione Data Mart	Data del caricamento del dato sul DM
Data elaborazione	ELAB/OUT	ETL di alimentazione Data Mart	Ultima elaborazione, anche se non terminata correttamente
Data presa in carico	ELAB/OUT	ETL di alimentazione Data Mart	Data di presa in carico del dato. Serve nel caso in cui il caricamento sui Data Marts non sia stato portato a termine ed abbia lasciato dei risultati parziali
Dati di auditing prestazionale	OUT	ETL di alimentazione Data Mart	Sono le prestazioni dell'ETL
Dati di auditing prestazionale in interrogazione	OUT	OLAP	Sono le prestazioni dell'OLAP
Destinatario del dato	IN/ELAB	ETL di alimentazione Data Mart	Area di analisi
Formato dei dati	IN	ETL di alimentazione Data Mart	Formati e dimensioni
Logs	OUT	ETL di alimentazione Data Mart	Log Applicativo generato dalla procedura di ETL in cui vengono memorizzati i singoli eventi ritenuti rilevanti ai fini delle successive fasi di tuning e controllo
Mappatura	ELAB/OUT	ETL di alimentazione Data Mart	Per singolo campo l'associazione con il campo dell'EDW
Periodicità di aggiornamento	IN	ETL di alimentazione Data Mart	Periodicità, orario, data di riferimento
Procedura di trasporto in ingresso	IN	ETL di alimentazione Data Mart	Nome della procedura
Regole di aggregazione e di calcolo	IN	ETL di alimentazione Data Mart	Sono regole per il trattamento tecnico dei dati
schema logico e fisico se si tratta di table db	IN	dbms	Nomi, formati, relazioni. Implicito nelle tabelle di sistema
Stato del dato sui DM	ELAB	ETL di alimentazione Data Mart	Completo, da integrare, da ricaricare. Da dettagliare sul singolo controllo quando si richiama il processo di aggregazione e le politiche di storicizzazione
Tipo dato/attributo	IN	ETL di alimentazione Data Mart	Dato elementare/aggregato/primitivo/derivato

Metadati di business

Metadato	IN/OUT/ELAB	Responsabilità del dato	Note
Aggregazione a cui partecipa	IN	ETL di alimentazione Data Mart	(vuote)
dato presente/aggiornato	ELAB/OUT	ETL di alimentazione Data Mart	Per tutti i dati di business di cui si vuole evidenziare la presenza. Ex. esiste il saldo contabile? É caricato ? É aggiornato e corretto?
definizione delle misure	IN	ETL di alimentazione Data Mart	(vuote)
Definizione gerarchie	IN	ETL di alimentazione Data Mart	(vuote)
KPI	IN	ETL di alimentazione Data Mart	(vuote)
Periodo di validità	ELAB/OUT	ETL di alimentazione Data Mart	(vuote)
Profili utenti e regole di visibilità	IN	ETL di alimentazione Data Mart	(vuote)
Regole di confronto e trattamento dei dati legati alla loro valore semantico e al significato di business	IN	ETL di alimentazione Data Mart	(vuote)
regole per la determinazione delle soglie di allarme	IN	ETL di alimentazione Data Mart	Sono regole per il trattamento dei dati per finalità di business.
Ritardo	ELAB/OUT	ETL di alimentazione Data Mart	con quale ritardo è entrato nel sistema
Storicizzato ?	IN	ETL di alimentazione Data Mart	(vuote)