Standard datawarehouse
Architettura

INDICE

1. Introduzione

Il presente documento fa parte della collezione di documenti "data warehouse", che raccoglie le indicazioni specifiche delle varie componenti sotto un unico indirizzo, separando al contempo le varie tematiche, al fine di proporre una più agevole lettura e meglio indirizzare le necessità di aggiornamento dei singoli documenti.

I documenti che compongono la collezione sono:

Architettura Data warehouse
E' il presente documento.

Indicazioni per lo sviluppo e realizzazione di progetti di Data warehouse
Il documento, rivolto in modo particolare ai responsabili dei progetti di Data Mart, intende rappresentare un ausilio concreto per coloro che partecipano alle singole iniziative progettuali del Ministero e non una linea guida teorica. Lo scopo è quello di affrontare lo sviluppo e la realizzazione delle inziative di natura informativa in corso presso il Ministero in modo integrato e uniforme, in coerenza con le scelte architetturali.

Il disegno fisico degli ambienti
Il documento descrive gli ambienti tecnologici previsti per ospitare sia l'Enterprise Data Warehouse sia i singoli progetti di Data Mart.

1.1 *Destinatari del documento*

Il documento è rivolto a tutti coloro che, operando nell'ambito del Ministero del Tesoro, del Bilancio e della Programmazione Economica, sono interessati alla definizione, alla progettazione e alla realizzazione di soluzioni di Data Warehouse.

1.2 *Scopo del documento*

Lo scopo del presente documento è quello di fornire un'architettura di riferimento, dettagliata nelle sue componenti principali, per i progetti di Data Warehouse.

Nel primo capitolo sono descritti i concetti base del Data Warehouse e sono sottolineate le differenze di obiettivi e caratteristiche rispetto ai sistemi di tipo gestionale.

Il documento è orientato soprattutto alla definizione di standard tecnologici. Tuttavia, in collaborazione con la Direzione Sviluppo Sistema Informativo e con la Direzione Nuovo Sistema Informativo, che sta al contempo affrontando la tematica da un punto di vista applicativo, si è ritenuto opportuno trattare anche gli aspetti di approccio metodologico legati allo sviluppo di progetti di Data Warehouse. Tali concetti sono riassunti nel capitolo 2.

Nel capitolo 3 è descritto uno schema di riferimento (framework), nell'ambito del quale si individuano le componenti tecnologiche e gli strumenti. Per le tre componenti (acquisizione di dati, gestione e distribuzione) sono descritti i requisiti e le funzionalità richieste. Si sono quindi indicati anche i prodotti di mercato che meglio rispondono alle esigenze e si integrano con il sistema informativo del Ministero.

Infine, il capitolo 4 riassume le indicazioni rispetto alle scelte tecnologiche adottate per il Ministero, che hanno lo scopo di rendere omogenee le soluzioni che si adotteranno per i vari progetti di Data Warehouse.

1.3 *Fonti*

Le scelte tecnologiche e metodologiche indicate nel presente documento tengono conto dei sistemi esistenti nei diversi Dipartimenti, che rappresentano dei vincoli in termini di operatività.
Si sono valutate inoltre le indicazioni tecnologiche e metodologiche dell'AIPA, di osservatori del mercato dell'Information Technology (Gartner Group, Ovum, Dataquest) e di società di consulenza (Technology Transfer).

2. Concetti preliminari di Data Warehouse

2.1 *Definizioni e obiettivi*

Tra le definizioni più diffusamente riconosciute di Data Warehouse, due in particolare ne identificano le caratteristiche peculiari:

"Una piattaforma sulla quale vengono archiviati e gestiti dati provenienti dalle diverse aree dell'organizzazione; tali dati sono aggiornati, integrati e consolidati dai sistemi di carattere operativo per supportare tutte le applicazioni di supporto alle decisioni " (Gartner Group)

"Un insieme di dati subject oriented, integrato, time variant, non volatile costruito per supportare il processo decisionale" (W.H.Inmon)

Si sottolinea quindi anzitutto la caratteristica del Data Warehouse come collezione di dati a supporto del processo decisionale del management.

Il Data Warehouse raggruppa i dati decisionali per aree o temi di interesse e li organizza rispetto all'utilizzazione finale; si differenzia in questo dai tradizionali data base il cui disegno è guidato dai requisiti delle applicazioni che garantiscono i processi gestionali.

In ambito Data Warehouse le informazioni assumono un valore aziendale piuttosto che dipartimentale, perchè sono il risultato integrato di dati provenienti da più fonti, anche esterne all'azienda. Tutte le informazioni sono rese coerenti da un modello comune dei dati e dalla definizione di standard aziendali (naming convention, unità di misura, codifiche).

Ha un orizzonte temporale ampio, garantendo il mantenimento di informazioni storiche, in modo da poter favorire le attività di analisi comparative su diversi periodi temporali. Le informazioni sono consolidate, consistenti nel tempo e non modificabili dall'utente che le accede esclusivamente in lettura.

Con il termine Data Mart si indicano delle collezioni di informazioni mirate ad un utenza dipartimentale e orientate ad un tema specifico. Il livello di aggregazione dei dati nel Data Mart risulta spesso più alto che nel Data Warehouse, che contiene anche dati di dettaglio, in quanto è disegnato per soddisfare in modo più diretto ed esplicito le esigenze dell'utente finale.

E' evidente l'importanza del principio della separazione tra ambienti operazionali e ambienti decisionali, informativi. La contemporanea insistenza di due classi di utenza di diversa natura e con diversa funzionalità sulla stessa base dati porrebbe problemi di contesa, creando tempi di risposta non pianificabili e spesso inaccettabili per i processi transazionali.
D'altronde la necessaria separazione degli ambienti, ottenuta attraverso un processo di mera replica degli ambienti transazionali non risponde comunque alle esigenze, perchè i database operazionali sono mirati al processo e non forniscono una visione integrata dei dati di interesse.

L'obiettivo del Data Warehouse è dare una risposta tempestiva e corretta a problemi decisionali legati alle attività istituzionali dell'Amministrazione, garantendo migliori risultati in termini di efficacia ed efficienza.

L'esigenza informativa ovviamente non è nuova, ma da un punto di vista tecnologico oggi esistono le condizioni per supportare questa richiesta in maniera adeguata. Gli aspetti tecnologici abilitanti sono la disponibilità di potenza elaborativa per gestire grossi volumi di dati a basso costo e la possibilità di distribuzione delle informazioni tramite tecnologia intranet/internet.

Va sottolineato però che, se il fattore tecnologico è presupposto necessario per lo sviluppo di soluzioni di Data Warehouse, il fattore critico di successo è il coinvolgimento dell'utente finale. Solo un'adeguata sponsorizzazione da parte del management può garantire il necessario contributo nel progetto di diversi settori aziendali con il relativo apporto di conoscenza dei processi e dei dati .

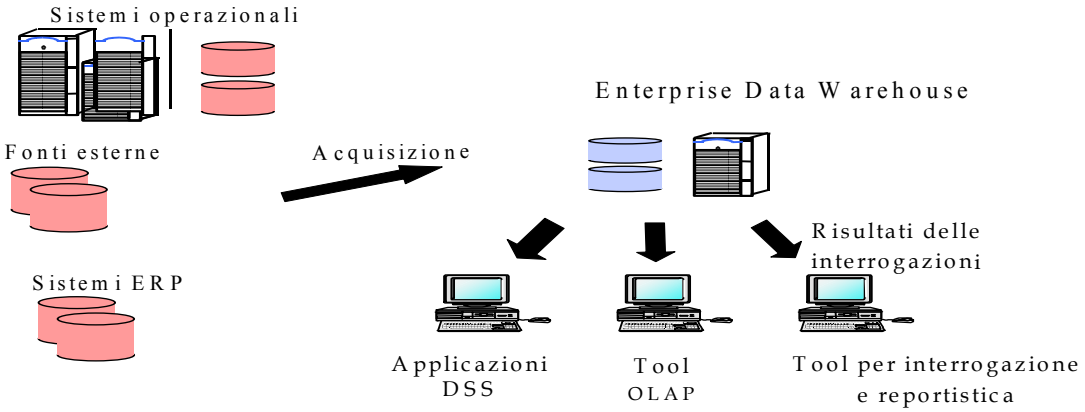
2.2 *Approccio metodologico al Data Warehouse*

L'approccio metodologico per la realizzazione di soluzioni di Data Warehouse è funzione dell'organizzazione aziendale, della tipologia di utenti, dello scopo che si intende perseguire e dell'architettura tecnica del sistema.
Nel corso del tempo sono stati suggeriti diversi approcci per la realizzazione dei progetti, poi rivisitati grazie al bagaglio di esperienze maturate da varie Organizzazioni. Si parla pertanto di approccio top-down, approccio bottom-up, approccio incrementale, a cui corrispondono diverse topologie di Data Warehouse (Enterprise Data Warehouse, Data Mart, Multi-tier Warehouse).

2.2.1 Approccio top down

L'approccio top-down è quello che prevede una implementazione estensiva del sistema, il cui disegno originale esamina fin dall'inizio tutte le principali aree di interesse aziendale. Si parla in questo caso di Enterprise Data Warehouse, che può essere successivamente suddiviso in un insieme di Data Mart, per motivi tecnici ed organizzativi. I Data Mart dipendenti costituiscono un subset di dati aziendali altamente specializzati per aree di interesse o dipartimenti aziendali.
Il punto debole di questo approccio teoricamente rigoroso è nella difficoltà di gestione del progetto omnicomprensivo, che rischia di paralizzare l'attività e di fornire risultati troppo in avanti nel tempo.

Enterprise Data Warehouse

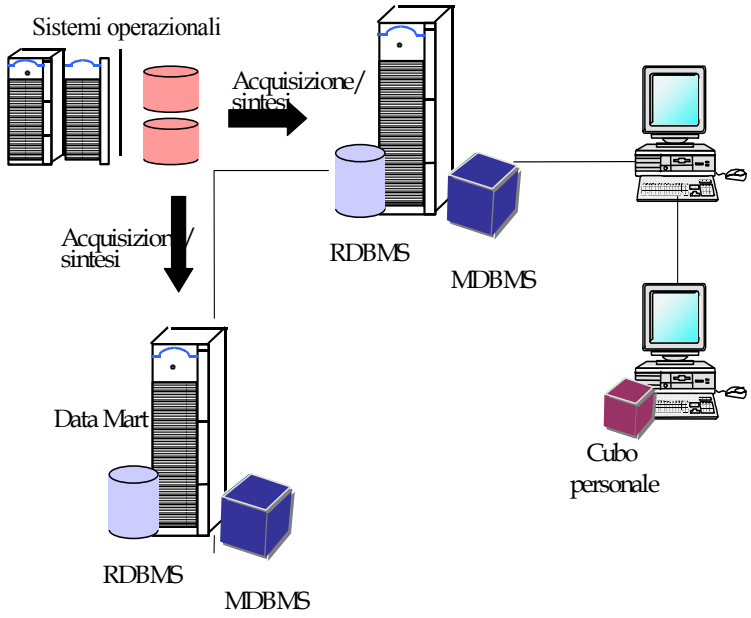


2.2.2 Approccio bottom-up

L'approccio bottom-up prevede una implementazione non coordinata nella quale ogni Data Mart viene realizzato per rispondere ad uno specifico fabbisogno informativo di una utenza dipartimentale. In questo caso l'Enterprise Data Warehouse è il risultato dell'insieme dei singoli Data Mart indipendenti, che si alimentano direttamente dai sistemi operazionali.

Il vantaggio di tale approccio pragmatico è conseguire risultati utili per l'utente in un arco temporale limitato con costi diretti relativamente contenuti. D'altra parte, mancando una visione iniziale complessiva, il rischio è quello di realizzare segmenti non integrabili fra loro, che originano isole informative probabilmente in parte ridondanti e non congruenti nei risultati.

Data Mart



2.2.3 Approccio "incrementale"

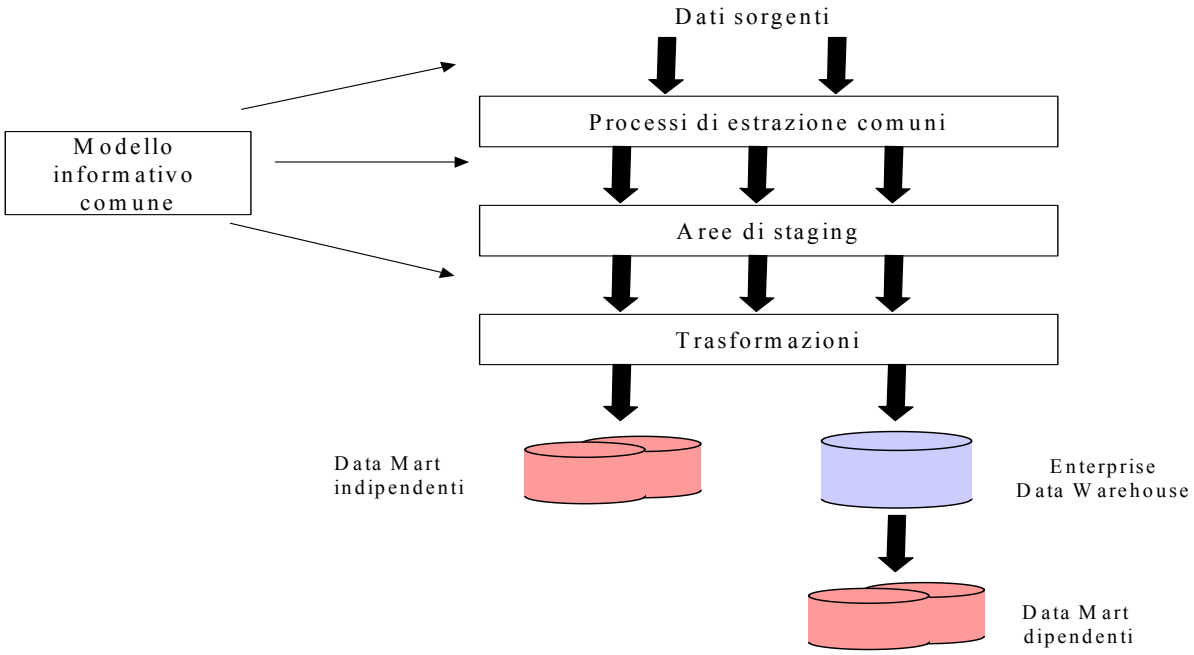
L'approccio "incrementale" combina i vantaggi dei due approcci sopra descritti. Alla base di questo approccio, definito in letteratura anche approccio "federato", infatti, è la creazione di un modello informativo comune.

Dal modello informativo comune vengono sviluppati in maniera coerente modelli dati dell'Enterprise Data Warehouse e/o dei Data Mart; questi ultimi possono essere sia dipendenti che indipendenti.

L'implementazione prevede di mettere a fattor comune tra diversi progetti di Data Mart i processi di acquisizione di dati dai sistemi source. Il risultato dei processi di acquisizione viene centralizzato su aree di appoggio comuni (cosiddette aree di staging) su cui vengono svolti i successivi processi di trasformazione. Le aree comuni di staging sono aree tecniche, non accedibili dall'utente finale, utilizzate per acquisire e trattare i dati con cui alimentare sia sia l'Enterprise Data Warehouse sia i Data Mart.

Il modello informativo comune e la fruizione delle aree di staging minimizza i problemi di integrazione tra Data Mart. L'implementazione delle soluzioni verso gli utenti risulta più rapida rispetto all'approccio top-down, perchè non è richiesto un modello dati enterprise disegnato completamente a priori, ma esso viene realizzato tramite un processo iterativo di definizione di aree tematiche di interesse prioritario. Ovviamente la gestione centralizzata di documentazione comune richiede architetture di sviluppo di tipo groupware e riduce in qualche modo l'autonomia (e l'anarchia) dei singoli gruppi.

Data Warehouse "Multi-tier"



Consip intende seguire l'approccio *incrementale* per assicurare un coerente sviluppo delle iniziative di Data Warehouse attualmente in corso e future.
Sarà garantita l'autonomia delle singole realizzazioni, assicurando al contempo la coerenza del significato semantico e della valorizzazione delle informazioni di utilizzo comune.

A tal fine sarà costruito un modello informativo comune, a partire dalla definizione dei processi e dei requisiti informativi espressi dagli utenti.
E' fondamentale, nell'approccio iterativo, che si definiscano le priorità per questi processi e si individuino delle aree tematiche.
Per ciascuna area tematica si analizzeranno i requisiti dati, le regole di business e di qualità.
Anche per la costruzione del modello informativo comune, quindi, si intende adottare un approccio incrementale. Esemplificando, alla costituzione del primo data mart, il suo modello dei dati e i relativi metadati coincideranno con il modello informativo comune; nel seguito i modelli provenienti dai successivi progetti verranno via via integrati nel modello informativo comune.
Il modello informativo comune potrà essere alimentato anche dalla documentazione dei dati provenienti dall'esterno e dalla documentazione, sotto forma di cataloghi/data dictionary,

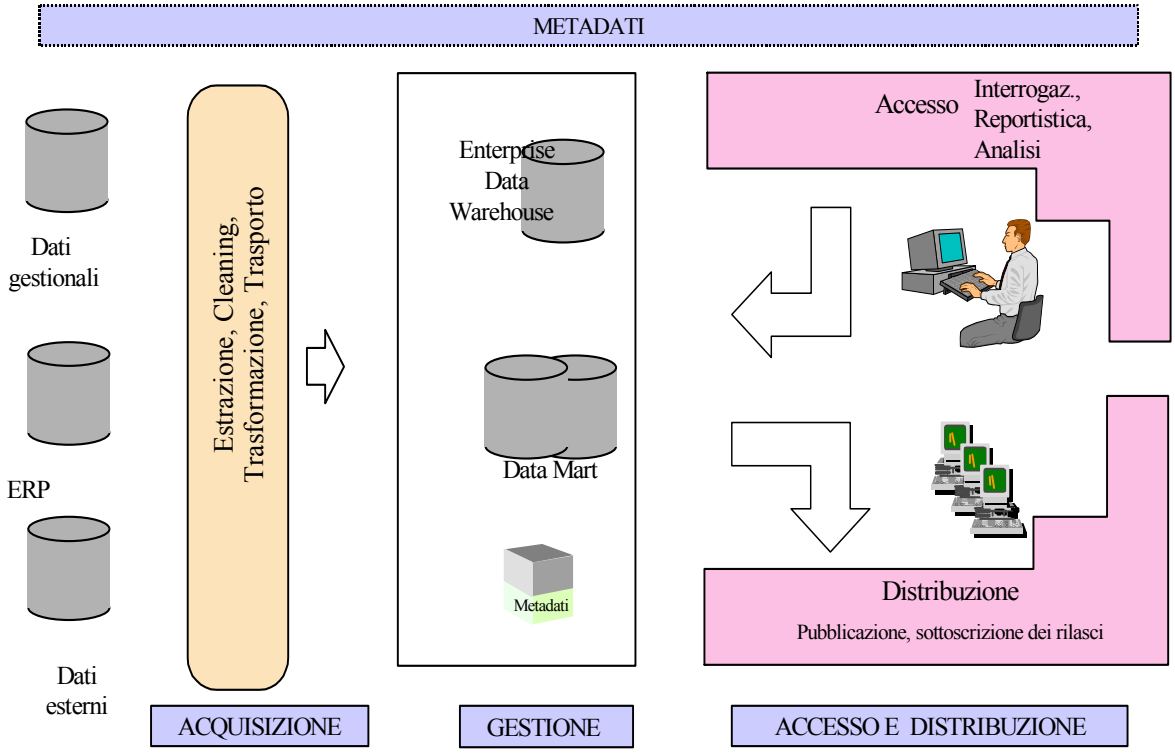
dei dati operazionali. Il recupero di tali informazioni richiederà una verifica dell'allineamento di quanto documentato nei cataloghi/data dictionary, con quanto effettivamente implementato a livello concettuale, logico e fisico.
Dal modello informativo comune si costruiranno il modello dell'Enterprise Data Warehouse e i modelli dei Data Mart, secondo le esigenze espresse dall'Amministrazione.
Tale modello potrà anche essere consultato dagli utenti per condividere conoscenze sul patrimonio informativo del Ministero.

Per tale attività occorre valutare prodotti di mercato che gestiscono metadati, supportando la documentazione di modelli dati e di processi e che consentano la acquisizione del patrimonio informativo attualmente gestito nel Ministero.

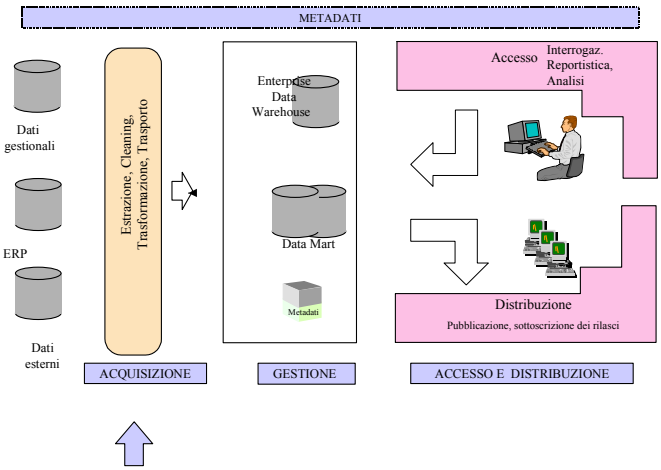
3. Il modello di riferimento

La definizione dell'architettura tecnica per progetti di Data Warehouse comporta la condivisione preliminare di un modello di riferimento che rappresenta le componenti funzionali necessarie allo sviluppo di progetti di Data Warehouse, aggregandole in tre aree: acquisizione, gestione, accesso e distribuzione.
Per ciascuna componente verrano indicati i principali requisiti ed una selezione di prodotti di mercato che rispondono alle esigenze. Alcuni prodotti sono stati scelti come standard Consip, altri sono in via di definizione.

Il modello di riferimento



3.1 *Acquisizione dati*



In questo paragrafo sono affrontati gli aspetti legati alle funzionalità previste nel processo di acquisizione dei dati di interesse del Data Warehouse dai diversi sistemi source. Sono quindi descritti i requisiti e indicato il tool di mercato selezionato per la soluzione da implementare presso il Ministero.

Il processo di acquisizione è normalmente la componente più costosa e complessa sia nella fase di sviluppo sia in quella di manutenzione del Data Warehouse. Per acquisizione si intende il processo nell'ambito del quale sono svolte le seguenti attività:

- **Estrazione dei dati dai sistemi source**
  Permette di acquisire dai sistemi operazionali e da fonti esterne i dati utili al processo decisionale. Dopo il primo caricamento, l'estrazione deve essere in grado di rilevare ed estrarre solo le variazioni dei dati nei database sorgenti.
  In alcuni casi, ad esempio per strutture con pochi dati e scarsa volatilità, per le quali l'informazione delle variazioni ha solo un'importanza storica, l'individuazione delle variazioni può essere demandata alle fasi di trasformazione e caricamento.
  A seconda dei requisiti delle singole iniziative, le estrazioni successive possono essere attivate dagli eventi di modifica oppure essere eseguite a determinati intervalli di tempo.

- **Cleaning**
  I dati estratti vengono raccolti in aree di staging e qui sottoposti a controllo di qualità tecnica e funzionale.
  La qualità tecnica riguarda soprattutto la presenza di codifiche disomogenee per dati che hanno lo stesso contenuto informativo ma provengono da sistemi diversi, o formati da convertire tra sistemi operativi diversi.

La qualità funzionale è notevolmente più complessa e coinvolge più direttamente l'utente finale del sistema di Data Warehouse in quanto è l'unico a poter certificarne la correttezza semantica.
Il processo di cleaning che mira alla omogeneizzazione dei dati, può evidenziare carenze e/o incompletezze che inducono un processo di correzione a livello dei dati sorgente.

- **Trasformazione**
  L'attività di trasformazione applica alle informazioni estratte e bonificate l'insieme di regole che li rendono rispondenti ai requisiti funzionali e tecnici richiesti dall'EDW e dai DM.
  Si eseguono pertanto attività di riorganizzazione e aggregazione di dati.
  I dati vengono inoltre arricchiti attraverso elaborazioni fatte rispetto alle diverse dimensioni di aggregazione previste dal disegno dei singoli Data Mart, in particolare per la caratterizzazione della dimensione temporale.

- **Caricamento**
  I dati estratti e trasformati vengono caricati sugli ambienti di Data Warehouse o Data Mart per l'accesso da parte degli utenti. La fase di caricamento può avvenire secondo differenti modalità di aggiornamento: inserimento o modifica.

3.1.1 Requisiti funzionali per i tool di acquisizione dati

I tool ETL (Extract, Trasformation and Loading), soprattutto a partire dalla metà degli anni '90, sono stati migliorati e arricchiti tanto da supportare il processo di acquisizione in modo semi-automatico.
I vantaggi collegati al loro utilizzo sono quelli di un più rapido sviluppo del codice, minori costi di manutenzione e il supporto alla definizione e gestione dei metadati tecnici.

I principali requisiti funzionali che sono valutati per i tool ETL riguardano:

- Capacità di acquisire dati da sorgenti diverse: RDBMS (in ambiente Unix e MVS), VSAM, sequenziali, sistemi ERP, Web
- Supporto alle aree di staging
- Capacità di avere un elevato numero di DBMS come possibili target
- Supporto alla verifica e alla "misurazione" della qualità dei dati
- Capacità di gestire i metadati tecnici e di business durante tutte le fasi del processo
- Capacità di definire e mantenere regole di cleaning e di trasformazione dei dati
- Possibilità di memorizzare i metadati tecnici in un database relazionale standard
- Possibilità di interrogare i metadati tecnici tramite strumenti di front-end standard
- Capacità di generare automaticamente procedure di estrazione, trasformazione e caricamento
- Possibilità di analisi di impatto sulle procedure di estrazione, trasformazione e caricamento rispetto alle variazioni apportate alle strutture dati source e/o target

- Possibilità di schedulazione e di journaling dei processi di estrazione, trasformazione e caricamento, sia dalla stazione di ammininistrazione, sia tramite lo schedulatore di sistema OPC
- Capacità di supportare processi di aggiornamento incrementale
- Capacità di documentare tutti i processi di trasformazione sui dati, tenendo traccia di eventuali variazioni
- Capacità di integrarsi con strumenti di modellazione dati per derivare le informazioni sugli archivi operazionali preesistenti e per alimentare i metadati con le informazioni descrittive del modello dati dell'EDW e dei Data Mart

Accanto ai suddetti requisiti funzionali, il tool ETL deve garantire semplicità di installazione e personalizzazione, facile gestibilità e flessibilità rispetto alla crescita dei volumi di dati e dei DBMS source e target.

3.1.2 Prodotti presenti sul mercato

I prodotti ETL di mercato si distinguono in due principali categorie di prodotti: i generatori di codice e i cosiddetti "transformation engine"

Fanno parte della categoria di generatori di codice, cioè dei prodotti ETL di prima generazione,:

- ETI-ETI Extract (partner IBM)
- Platinum Technology- DecisionBase
- Prism - Prism Executive Suite (acquisita da Ardent)
- SAS- Warehouse Manager

Fanno parte della categoria "transformation engine", cioè dei prodotti ETL di seconda generazione,:

- Ardent -Data Stage
- Carleton- Passport ed Enterprise Integrator
- Constellar- Constellar Hub e Warehouse Builder
- Informatica - PowerMart e PowerCenter
- CA -Information Builders

3.1.3 Considerazioni sull'acquisizione dei dati

In considerazione del numero di iniziative di Data Warehouse attualmente in corso al Ministero e in previsione di future esigenze, Consip reputa conveniente realizzare i processi di acquisizione facendo ricorso ad un tool di mercato piuttosto che procedere con sviluppo di SW ad hoc.
.
Tra i tool citati Consip ha individuato il prodotto Informatica Power Center, della categoria "transformation engine", come il più adatto a rispondere ai requisiti.

Nella scelta si è tenuto conto dei nuovi indirizzi del mercato, avvalendosi dei risultati che emergono dalle analisi di Gartner Group e Ovum.
Si è anche proceduto alla sperimentazione presso il CED di La Rustica su dati di un'applicazione campione del Ministero.

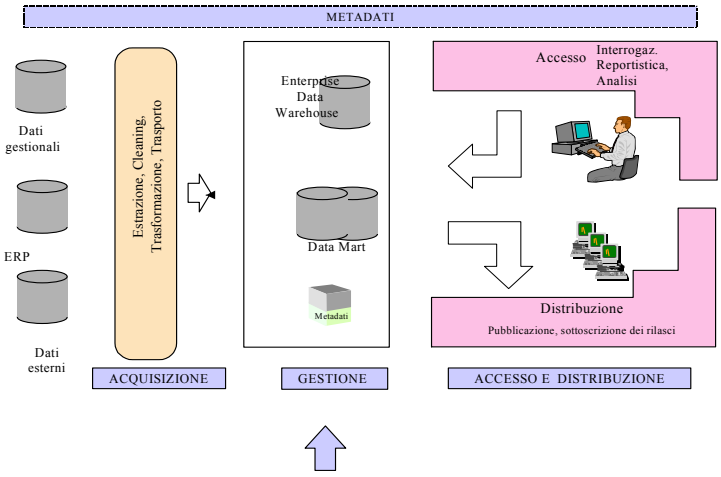
Power Center appartiene alla categoria di prodotti ETL cosiddetti "trasformation engine".
Questi prodotti, a differenza dei tool ETL di prima generazione, non generano codice ma creano automaticamente delle procedure che eseguono l'intero processo di estrazione, trasformazione e caricamento, semplificando notevolmente le operazioni di gestione.
L'orientamento del mercato è verso i trasformation engine, come risulta anche dalle indicazioni di tutti gli osservatori.

Power Center consente di documentare in maniera centralizzata, omogenea e riutilizzabile i cosiddetti metadati tecnici, che si originano dai processi di estrazione, trasformazione e caricamento dei dati dagli ambienti legacy a quelli decisionali.
Il prodotto garantisce la più completa integrabilità con altre componenti architetturali: supporta le principali piattaforme hardware e software, consente la schedulazione delle procedure create anche tramite schedulatore di sistema OPC, consente di memorizzare il suo repository su un DBMS relazionale standard (Oracle) e di accedere ai metadati tramite strumenti di query e reporting di mercato (ad esempio Business Objects).
Il tool consente, inoltre, di eseguire in modo automatico l'analisi di impatto legata alle variazioni delle strutture dati in input o in output, permettendo la modifica delle procedure di estrazione, trasformazione e caricamento senza richiedere una costosa manutenzione del SW sviluppato.

Per logiche di trasformazione particolarmente complesse e non supportate dal tool, è possibile scrivere delle exit routine, che il tool è in grado di acquisire e documentare nel repository.

Pur non avendo funzionalità specifiche legate alla "pulizia" dei dati, il tool Power Center consente di valutare il grado di qualità dei dati attraverso l'applicazione di condizioni di filtro e di migliorarne il livello attraverso processi iterativi di elaborazione.
Sono allo studio anche tool specifici legati alla qualità dei dati, ma si ritiene opportuno valutarne l'impiego solo dopo l'esame di quanto emergerà dai primi processi di estrazione e trasformazione realizzati con Power Center.

3.2 *Gestione dati*

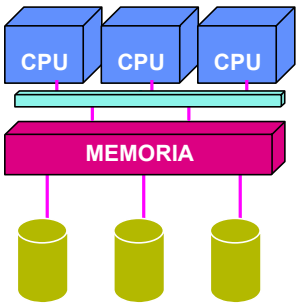


In questo paragrafo si esamina la gestione dati dei sistemi di Data Warehouse, sia in termini di modellazione sia in termini di infrastrutture tecnologiche.
Si valutano le piattaforme HW, i Sistemi Operativi e i DBMS adeguati alla gestione dell'Enterprise Data Warehouse, dei Data Mart e delle aree di staging. Nell'esaminare singolarmente le componenti vengono fatte alcune semplificazioni, ma di fatto l'architettura complessiva è il risultato dell'interazione delle tre componenti. Questo comporta una soluzione pratica con caratteristiche ibride che possono evidenziare i vari aspetti delle diverse opzioni tecnologiche.

3.2.1 Piattaforme HW

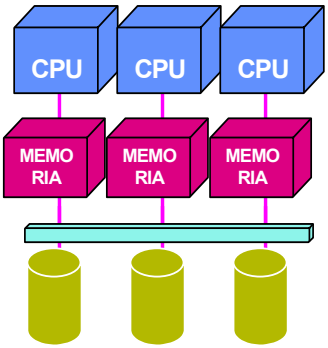
Le soluzioni HW più rispondenti alle funzioni richieste dai sistemi di Enterprise Data Warehouse e per Data Mart sono quelle fornite da macchine con tecnologia parallela. In questa tipologia oggi rientrano quattro tipi di architetture, di seguito riportate, che si distinguono per le modalità di interazione fra i tre elementi base: processori (CPU), memoria, dischi.

- Sistemi SMP (Symmetric Multi Processing)



I sistemi SMP sono caratterizzati da un'architettura in cui esistono un numero variabile di CPU, che condividono la memoria ed i dischi. In sistemi di questo tipo la memoria è l'elemento condiviso principale e le CPU devono gestirne l'accesso concorrente.
Questo tipo di sistemi, pur garantendo prestazioni significative rispetto a volumi significativi di dati, trovano il loro limite nella condivisione delle risorse. Oltre certi livelli prestazionali, l'incremento della potenza elaborativa non comporta eguali benefici in termini di performance, dato il collo di bottiglia sugli elementi condivisi del sistema.

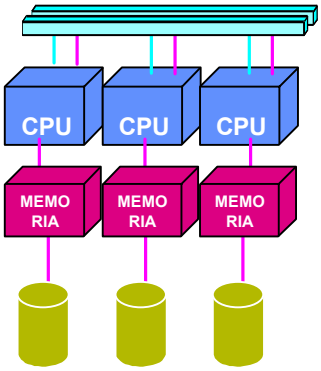
- Sistemi Cluster-SMP (Shared Disks)



I sistemi Cluster-SMP (o semplicemente Cluster) sono quei sistemi caratterizzati dall'avere un numero variabile di CPU ognuna delle quali è connessa ad un modulo di memoria privato. In sistemi di questo tipo, i dispositivi di memoria di massa (dischi), essendo

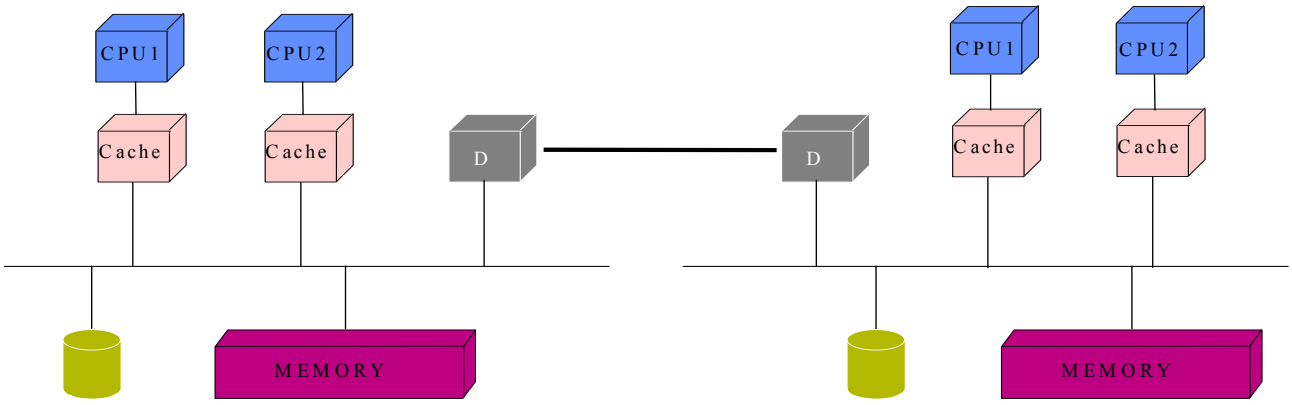
condivisi, rappresentano un potenziale collo di bottiglia che può penalizzare le performance globali del sistema. Il collegamento tra le varie CPU è solitamente ottenuto attraverso LAN ad alta velocità.

- Sistemi MPP (Massive Parallel Processing)



Sono anche detti sistemi Shared Nothing (nessuna condivisione). La loro caratteristica è quella di avere sottosistemi autonomi (nodi) dotati di CPU, memoria e dischi propri, collegati tra di loro attraverso un meccanismo di interconnessione proprietario ad altissima velocità ed affidabilità. Questi sistemi hanno una scalabilità di prestazioni quasi lineare e sono particolarmente indicati per la gestione di Enterprise Data Warehouse e Data Mart caratterizzati da elevatissimi volumi dati (centinaia di Gigabyte ed oltre). Di contro la gestione di sistemi MPP risulta essere più costosa e più legata all'offerta di pochi fornitori rispetto ai precedenti sistemi.

- Sistemi NUMA (NonUniform Memory Access)
  I sistemi NUMA sono costituiti da una serie di sistemi SMP le cui memorie sono connesse mediante un bus ad altissima velocità. Un esempio di un sistema NUMA costituito da due SMP è il seguente:



Il sistema mostrato nella figura precedente è caratterizzato da una connessione fra le due componenti denominate **D.** Tali componenti ridirigono le richieste di memoria in maniera appropriata mediante un'interconnessione con l'altro nodo. In particolare, quando un processore richiede l'accesso ad una certa zona di memoria, la componente D si preoccupa di rendere completamente trasparente l'indirizzamento occupandosi anche di risolvere eventuali problemi di coerenza del dato fra le diverse memorie (CC_NUMA).
Un sistema di questo tipo possiede, a differenza dei sistemi SMP in cluster o MPP, un unico sistema operativo. Inoltre non è soggetto alle difficoltà insite nelle attività di amministrazione delle altre architetture. Ovviamente, essendo il sistema operativo unico, l'eventuale indisponibilità di un processore può comportare l'indisponibilità dell'intero sistema. In questo senso esistono delle modifiche architetturali, (es. SUN ULTRA 10000) che permettono di partizionare il sistema in sistemi logici separati consentendo, quindi, di indirizzare problematiche di alta disponibilità del sistema.

3.2.1.1 Requisiti per le piattaforme hardware

Le piattaforme hardware da utilizzare per soluzioni di Data Warehouse devono assicurare:

- Scalabilità lineare
  Per scalabilità lineare si intende la capacità di supportare un numero doppio di utenti raddoppiando le risorse della macchina. Ovviamente, qualunque sia l'architettura della macchina, è molto complicato sia realizzare una reale

duplicazione di tutte le risorse coinvolte nel processo elaborativo, sia misurare e comparare i diversi tempi di risposta con i benchmark attualmente a disposizione. In ogni caso l'architettura MPP è quella che maggiormente indirizza tale aspetto in quanto non presenta colli di bottiglia architetturali

- Flessibilità ed espandibilità
  L'architettura deve avvalersi di una tecnologia tale da permettere ampi incrementi di potenza senza il ricorso a pesanti modifiche nella configurazione della macchina. Deve, inoltre, essere possibile far coesistere tipologie diverse di configurazione in maniera tale da indirizzare con maggiore efficacia la gestione di più ambienti operativi (sviluppo, test, collaudo) e di esercizio (batch, datamart,...). In particolare la tecnologia MPP o SMP clustering potrà avvalersi dell'isolamento architetturale che esiste fra i nodi del sistema, consentendo una diversificazione in termini di CPU, RAM, disco e adattatori. Nel caso, invece, di un'architettura diversa (SMP-NUMA) dovrà essere possibile implementare un meccanismo di partitioning in maniera tale da dividere il sistema in insiemi logicamente separati

- Affidabilità e disponibilità
  L'architettura hardware di riferimento deve permettere una facile implementazione di funzioni che assicurino alta disponibilità al sistema in termini di affidabilità dei dati e disponibilità del servizio. In questo senso tutte le architetture presentate, ad eccezione di SMP, si presentano in maniera adeguata. Ovviamente l'implementazione architetturale dovrà essere accompagnata da un opportuno software che consentirà una gestione evoluta delle eventuali malfunzioni relative alle diverse componenti.
  I sistemi potranno prevedere la duplicazione delle componenti hardware più critiche ed implementare delle politiche di gestione delle memorie di massa che consentano di sopperire ad eventuali guasti dei dischi senza interrompere il servizio, a seconda delle necessità e dei requisiti di affidabilità richiesti dal Ministero sulla singola iniziativa. Le componenti o i sistemi più critici dovranno essere ridondati in modo da assicurare la continuità del servizio in caso di indisponibilità del sistema principale. A tal fine sarà opportuno utilizzare software di Clustering, che permetta una gestione globale delle malfunzioni, dal guasto hardware alla caduta di un sottosistema software fino al ripristino delle funzionalità applicative

- Connettività
  E' un aspetto molto importante di un sistema Data Warehouse, soprattutto quando la quantità di dati in input è rilevante ed in continua espansione. In questo senso è fondamentale l'utilizzo di canali di comunicazione efficienti verso i sistemi gestionali di input. Nel caso, ad esempio, di un sistema MVS potrà essere valutato l'acquisto di canali ESCON diretti verso il sistema DW o, alternativamente, l'utilizzo di unità dischi SCSI connesse direttamente ai sistemi gestionale e DW con meccanismi automatici di conversione dati. E' importante, inoltre, garantire buone performance anche ai flussi di output verso i Data Mart, posizionando questi ultimi in maniera ottimale rispetto al DW centrale

3.2.2 Sistemi operativi

3.2.2.1 Requisiti per il sistema operativo

I principali requisiti previsti per il sistema operativo riguardano la

- Affidabilità
  Il S.O. deve essere totalmente integrato con la piattaforma hardware al fine di assicurare alti livelli di servizio per l'utente. Deve essere possibile sopperire a malfunzionamenti dovuti ad avarie nel software attraverso utility che consentano sia di monitorare eventuali degradi e/o malfunzionamenti sia di implementare delle funzioni automatiche di risposta coordinando le azioni di recovery

- Scalabilità
  Il S.O. deve supportare la più ampia scalabilità delle piattaforme hardware su cui risiede, dal desktop al supercomputer. E' fondamentale, date le caratteristiche in termini di espandibilità e scalabilità del sistema Data Warehouse, che vengano mantenute, al crescere della macchina ed al variare delle configurazioni, la portabilità e le performance dell'applicazione

- Supporto dell'architettura a 64 bit
  Il S.O. deve supportare processori a 64 bit, consentendo di indirizzare memorie e quindi cache di elevatissime dimensioni. Questo, se accompagnato da analoghe caratteristiche del database, permette di migliorare gli accessi alla base-dati del sistema DW

- Maturità
  Il S.O. deve presentare caratteristiche di maturità e consolidamento sul mercato con il pieno supporto delle nuove tecnologie. Questo consente di minimizzare gli impatti nella migrazione a nuove versioni

- Facilità di gestione
  Deve essere presente un software grafico evoluto, possibilmente Web, che consenta agli amministratori di gestire il sistema operativo e la piattaforma hardware in maniera efficiente; uno strumento, quindi, integrato con il S.O., che supporti "una grafica object oriented" e che semplifichi le attività operative sul sistema quali la gestione degli spazi e la configurazione delle interfacce di comunicazione

- Aderenza a standard pubblicati
  Nella scelta del sistema operativo, dovrà essere valutata la disponibilità di adeguamento dei prodotti software alle indicazioni architetturali di organismi di standardizzazione (X/Open XPG4, POSIX)

In generale, la scelta di un sistema per il Data Warehouse potrebbe essere indirizzata verso OS/390 o sistemi UNIX.
Nel primo caso si manterrebbe un forte vincolo con il fornitore, sia per quanto riguarda la piattaforma hardware sia per il DBMS dando, quindi, alla soluzione una connotazione di sistema "proprietario". Inoltre, i costi potrebbero essere di gran lunga superiori soprattutto per progetti caratterizzati da un numero di utenti e volumi di dati non particolarmente elevati. I vantaggi, al contrario, emergerebbero nel momento in cui le dimensioni previste fossero di un ordine di grandezza tale (Base dati > 500GB) da giustificare l'investimento.

I requisiti espressi non sono altrettanto stringenti nel caso di progetti di Data Mart di piccole dimensioni, per i quali non è prevista una estrensione significativa nel tempo né in termini di volumi di dati da gestire né in termini di utenza. Windows/NT alla data non è invece ancora sufficientemente maturo e stabile per soluzioni Enterprise o per Data Mart di cui si prevede uno sviluppo significativo nel tempo. Non risponde infatti ai requisiti di scalabilità ed affidabilità richiesti.

La valutazione complessiva della rispondenza ai requisiti tecnici, delle dimensioni dei Progetti da supportare e degli aspetti economici porta a scegliere la piattaforma Unix come quella più adatta per implementare un ambiente Enterprise Data Warehouse presso il Ministero del Tesoro.
I sistemi Unix sono peraltro già presenti presso diverse realtà dipartimentali.

3.2.2.2 Le soluzioni di mercato

Ovviamente la scelta Unix, in quanto non proprietaria, non identifica in maniera univoca una piattaforma hardware. Attualmente, sulla base dei requisiti indicati, possono essere identificati almeno 5 fornitori:

1. IBM
   - Sistema RISC SP2
     Tale macchina presenta un'architettura MPP ed il sistema operativo AIX V4.3. E' altamente scalabile ed espandibile con una rilevante diffusione sul mercato (circa 7000 nodi installati nel mondo). E' adatta a sistemi DW di oltre 150 GBSistema
   - S70
     Tale macchina presenta un'architettura SMP ed il sistema operativo AIX V4.3. Scalabile ed espandibile, è adatta a sistemi datawarehouse < 150GB

2. SUN
   - Sistemi Starfire 6000 e 10000
     Sono macchine con architettura NUMA e sistema operativo Sun Solaris V2.6, altamente espandibili. In particolare il sistema 10000 può avere fino a 64 processori con la possibilità di partitioning

3. HP
   - Sistemi HP9000 V2500 con Sistema Operativo HP-UX 11.00

4. DIGITAL
   - Sistemi della famiglia 8400, con architettura SMP clusterizzabile. Il sistema operativo è Digital Unix

5. SEQUENT
   - Sistema NUMA-Q 2000.
     E' una macchina con architettura NUMA e sistema operativo Unix-Dynix

I sistemi suddetti rispondono tutti ai requisiti di un sistema Data Warehouse.
Le differenze principali sono relative alla diffusione sul mercato, alla disponibilità sulla rispettiva piattaforma di prodotti di gestione dati ed al prezzo.
IBM, SUN ed HP presentano delle piattaforme Unix molto diffuse sul mercato, tecnologicamente molto avanzate e con ampia disponibilità di prodotti a supporto.
Sequent risente, al contrario, di una scarsa diffusione sul mercato, con possibili impatti anche sulla scelta dei pacchetti applicativi.
IBM e SUN dispongono dell'ulteriore vantaggio di essere già presenti presso alcune realtà del Ministero quali le RPS ed il III Dipartimento.

3.2.3 DBMS

Il DBMS rappresenta un elemento chiave nel disegno e realizzazione dell'intera soluzione Data Warehouse. Nella scelta del DBMS vengono considerati requisiti tecnici di operabilità in ambienti eterogenei. Viene valutata, inoltre, la diffusione del prodotto sul mercato.

3.2.3.1 Requisiti per il DBMS

Il DBMS a supporto dei progetti di Data Warehouse prevede i seguenti requisiti:

- deve avvalersi di una tecnologia relazionale
- deve soddisfare opportuni requisiti di standardizzazione, diffusione sul mercato e apertura verso la maggioranza dei pacchetti applicativi
- deve essere multi-piattaforma, al fine di rendere possibile l'utilizzo nativo delle funzioni sia in ambienti Windows/NT che in ambienti Unix e di garantire una completa portabilità delle strutture tabellari in ambienti operativi diversi
- deve fornire funzioni evolute di system management che consentano di gestire l'ambiente database da un unico punto di controllo, utilizzando un'interfaccia utente grafica. Inoltre devono essere disponibili funzioni di navigazione fra i vari oggetti del database, di schedulazione e di gestione automatica degli eventi
- deve indirizzare in maniera evoluta le problematiche di sicurezza dati ed utenti. In particolare deve essere possibile l'autenticazione dell'utente a livello interno e tramite interfaccia a pacchetti di sicurezza esterni (ad esempio, Dce, Kerberos)
- deve essere perfettamente integrato con la piattaforma hardware, consentendo il parallelismo e la distribuzione dell'elaborazione dati su più processori e/o nodi
- deve rispettare lo standard SQL-ANSI (Structured Query Language) per l'interrogazione e la manipolazione della base dati

- deve supportare le estensioni standard del linguaggio SQL di uso comune in un sistema DW con funzioni quali, ad esempio, il partitioning, indici bitmap, star query e hash join, materializzazione delle viste
- deve fornire un'integrazione completa con l'ambiente WEB sia a livello di supporto dell'architettura applicativa a 3 livelli che di predisposizione dell'ambiente verso tecnologie internet. In questo senso il database deve essere predisposto per supportare colloqui applicativi con altri sistemi in ottica RUPA
- deve fornire il supporto, in aggiunta agli alfanumerici, di altri tipi di dati: testuali, video, grafici, temporali ed immagini
- deve consentire l'utilizzo di tecniche di Triggers e Stored Procedure
- deve rendere disponibili funzioni per effettuare auditing affidabili e flessibili con la possibilità di registrare tutte le operazioni di database interessate
- deve fornire funzioni per effettuare salvataggi totali, incrementali, e on-line

3.2.4 L'architettura tecnologica per le aree di staging

Nell'ambito di grossi sistemi datawarehouse, assumono importanza le aree di staging.

Queste aree si configurano come aree tecniche, dove vengono consolidati ed aggregati gli output dei processi gestionali nella fase che precede l'alimentazione della base dati del Data Warehouse e dei Data Mart. Tale area dati risiede, in genere, sullo stesso sistema DW oppure su un sistema sorgente quale un sistema host, a seconda di dove risulti più conveniente utilizzare delle procedure batch per il "cleaning" e la trasformazione dei dati.
I dati delle aree di staging possono essere strutturati in DBMS oppure mantenuti sotto forma di file sequenziali.

3.2.5 Considerazioni sulla gestione dei dati

Sulla base della rispondenza ai requisiti descritti e in funzione della diffusione di sistemi Unix presso alcune realtà dipartimentali significative, la piattaforma Unix è quella più adatta per implementare un ambiente Enterprise Data Warehouse presso il Ministero del Tesoro.
Windows /NT può considerarsi adeguato per soluzioni specifiche di Data Mart, soprattutto nel caso in cui il numero di utenti rimanga circoscritto ad una decina e il volume complessivo di dati da gestire sia inferiore a 10 GB. Questo consente l'utilizzo di macchine a più basso costo e di più facile gestione.
Windows/NT alla data non è invece ancora sufficientemente maturo e stabile per soluzioni Enterprise o per Data Mart di cui si prevede uno sviluppo significativo nel tempo. Non risponde infatti ai requisiti di scalabilità ed affidabilità richiesti.

Sulla base dei requisiti sopra esposti è stato individuato il database Oracle quale standard per la realizzazione di soluzioni a supporto decisionale. Il database è adeguato sia per supportare la gestione dell'Enterprise Data Warehouse, sia per la gestione dei Data Mart.

Esso fornisce infatti tutte le funzionalità tipiche dei DBMS relazionali (back-up, recovery, gestione del locking, utility di gestione,..) ma anche aspetti tipici per il supporto agli ambienti decisionali. In particolare prevede dei meccanismi di partizionamento per la parallelizzazione delle operazioni e meccanismi di hash join. Inoltre, nel caso di query che si ripetono frequentemente, sulla base delle informazioni registrate nel log, è in grado di suggerire al DataBase Administrator la creazione di viste materializzate che utilizzerà automaticamente.
Utilizza lo standard SQL-ANSI per l'interrogazione e la manipolazione dei dati ed offre una completa integrazione e portabilità in ambienti eterogenei.
Inoltre il database si integra in modo coerente con un insieme di tool che permettono di generare applicazioni in modo agevole.
Tale scelta è stata fatta alla luce della presenza ormai consolidata di tale RDBMS presso alcune realtà dipartimentali del Ministero del Tesoro. Infine si è tenuto conto del fatto che diversi osservatori di mercato dell'IT (Gartner Group, Dataquest) lo indicano come leader assoluto nel mercato dei DBMS relazionali, in particolare sulle piattaforme Unix.

3.2.6 I metadati

Al crescere delle informazioni memorizzate nel Data Warehouse e all'aumentare delle modalità con cui queste vengono utilizzate, è necessario gestire in una sorta di repository i cosiddetti metadati, cioè le informazioni relative ai dati stessi.

La definizione di metadati varia in funzione del contesto in cui si applica:
- nel disegno del Data Warehouse rappresentano la mappatura delle informazioni di business sui dati contenuti nel Data Warehouse
- nell'acquisizione tramite tool ETL rappresentano le modalità di trasformazione dei dati dai sistemi operazionali al Data Warehouse
- nella gestione tramite DBMS rappresentano gli oggetti del database (es. tabelle, viste, utenti, ...)
- nell'accesso tramite prodotti OLAP rappresentano la mappatura dello schema fisico del database rispetto alla vista ottenibile tramite le query

I metadati vengono creati e gestiti durante la fase di disegno e sviluppo del Data Warehouse. Possono essere importati da sorgenti esterne quali i cataloghi del DBMS, librerie di programmi, prodotti Case.
Vengono gestiti all'interno dei tool che coprono i diversi aspetti architetturali. Spesso l'architettura dei tool è proprietaria e comporta una gestione non unitaria dei metadati.

Normalmente si distinguono in metadati tecnici e metadati di business o funzionali.

I metadati tecnici contengono informazioni dettagliate sulle fasi di disegno, sviluppo, creazione e gestione (autorizzazioni, frequenza dei salvataggi, frequenza degli aggiornamenti, versioning) dei dati del Data Warehouse. Altri esempi di metadati tecnici sono le informazioni legate all'acquisizione dei dati (regole di pulizia ed estrazione), le informazioni sulle tipologie di accesso al Data Warehouse e sul tipo di utilizzo dei dati che viene fatto da parte degli utenti (statistiche).

I metadati di business contengono informazioni che permettono all'utente finale di accedere al Data Warehouse in modo comprensibile dal punto di vista del business. Queste informazioni riguardano in particolare l'associazione tra metadati tecnici e concetti di business: da quali sistemi sorgente provengono i dati, dettagli relativi alle query, report e oggetti DSS predefiniti, sottoscrizioni a report e analisi i cui risultati vengono poi forniti con regolarità.

Il requisito fondamentale richiesto ai prodotti che supportano l'utilizzo di metadati è l'integrazione della loro gestione durante tutto il processo di creazione del Data Warehouse.

Ancora oggi però non esiste uno standard globale di gestione delle tipologie di metadati, che vengono creati e gestiti in maniera proprietaria all'interno dei tool che coprono i diversi aspetti architetturali.
La reale difficoltà è quella quindi di integrare e sincronizzare tali isole di metadati, per avere un'unica interfaccia e modalità di trattamento dei metadati tecnici e di business.

A tal fine si sono avviate alleanze tra fornitori di tool ETL e fornitori di tool di accesso, che però non hanno ancora portato a risultati concreti.

In assenza di un prodotto unico di mercato per la gestione integrata dei metadati tecnici e di business, Consip gestirà i metadati di business attraverso il tool di analisi e disegno (probabilmente il Kit dei prodotti della Sterling Software o ERwin) e i tool di accesso che mascherano lo strato fisico con la definizione di oggetti di business (es. gli universi di Business Objects).
I metadati tecnici verranno creati e memorizzati attraverso il tool ETL.

3.3 *La modellazione dei dati*

Le indicazioni fornite in questo paragrafo, lungi dall'essere esaustive, hanno il solo scopo di sottolineare le specificità dei progetti di Data Warehouse anche rispetto alle tecniche di modellazione dei dati.
Per maggiori approfondimenti, relativi alla modellazione dei dati, sia a livello di modello Enterprise sia a livello di Data Mart, si rimanda ai noti trattati di modellazione dimensionale (Ralph Kimball -"The Data Warehouse Lifecycle Toolkit"- Wiley- 1998;
W.H. Inmon - "Building the Data Warehouse"- Wiley- 1996)

La tipicità dei progetti di Data Warehouse, rispetto alle tradizionali applicazioni gestionali, si riflette anche in una diversa tecnica di modellazione dei dati.

In ambiente decisionale i dati vengono organizzati in modo da rispondere direttamente alle interrogazioni tipiche dell'utente finale.

In ambito Data Warehouse si parla quindi di modellazione dimensionale dei dati.

Il modello dimensionale e quello tradizionale entity-relationship sono potenzialmente in grado di memorizzare le stesse informazioni. La differenza è legata alla ricerca di prestazioni ottimali sulle due tipologie di ambienti.
Negli ambienti operazionali è fondamentale adottare tecniche di normalizzazione che, eliminando le ridondanze, permettono ad un elevato numero di transazioni ripetitive di interrogare ed aggiornare le informazioni in un singolo punto del database.
I sistemi di supporto alle decisioni, invece, sono caratterizzati da un ridotto numero di transazioni, che operano però in maniera trasversale sulla base dati, richiedendo l'accesso in lettura su un elevato numero di dati da mettere in join. L'uso di tecniche di normalizzazione appare quindi controproducente nella progettazione di un sistema di Data Warehouse. In ambiente decisionale è più corretto disegnare il database in modo da soddisfare in maniera diretta e semplificata le esigenze specifiche degli utenti finali, anche attraverso dati denormalizzati e aggregati.

E' importante dedicare molta attenzione all'analisi del modello concettuale, prevedendo tutte le dimensioni di analisi e per ciascuna dimensione i relativi attributi. Gli attributi delle tabelle dimensionali vengono usati direttamente come condizioni di ricerca nelle interrogazioni sul Data Warehouse e come intestazioni delle colonne nei report degli utenti finali. L'uso di dati preaggregati e di sintesi limita il numero di operazioni necessarie a ricostruire l'informazione richiesta.

L'implementazione tipica del modello dimensionale è quella che risulta nello schema dati a stella o star schema.

Lo star schema, oltre a essere ottimizzato per le interrogazioni, presenta l'ulteriore vantaggio di consentire una rappresentazione grafica del business più comprensibile per l'utente finale.

A differenza dello schema entità-relazioni, in cui tutte le entità sono rappresentate nello stesso modo, lo star schema è asimmetrico. Al centro del diagramma c'è una tabella dominante di grandi dimensioni, che è l'unica a possedere collegamenti multipli, che si concretizzano in campi chiave in essa contenuti, alle altre tabelle. Le altre tabelle rappresentano le dimensioni di analisi attraverso cui si esamina il "fatto" di interesse.
Tecnicamente si definiscono "fact table" la tabella centrale e "dimension table" le altre.
La fact table è la tabella in cui vengono memorizzate le misure numeriche del fatto (nell'esempio l'importo del finanziamento del progetto di investimento), ognuna delle quali rappresenta l'intersezione di tutte le dimensioni e quantifica una o più variabili. Normalmente è l'unica tabella normalizzata.
Le altre tabelle rappresentano le dimensioni di analisi attraverso cui si esamina il "fatto" e contengono attributi di tipo descrittivo. La cardinalità di queste tabelle, normalmente denormalizzate, è limitata.
Tra le tabelle dimensionali, la tabella della dimensione tempo è sempre presente nei modelli dimensionali, quale che sia il business da rappresentare. Questo deriva dalla caratteristica dei sistemi di Data Warehouse che gestiscono anche le informazioni storiche.
A volte, per migliorare ulteriormente la interpretabilità del modello, anche le tabelle dimensionali vengono normalizzate, esplicitando le gerarchie. Questo dà luogo ad uno schema detto a "fiocco di neve" o Snowflake Schema.

La navigazione dei dati attraverso il modello snowflake è meno efficiente che nello Star Schema perchè aumenta il numero di tabelle su cui effettuare le join.

Va considerato che, se da un lato la denormalizzazione garantisce migliori prestazioni nelle operazioni di interrogazione, la stessa rende più oneroso l'aggiornamento e il caricamento dei dati. Più in generale, le operazioni di manutenzione del modello dati multidimensionale, quali l'estensione a nuove dimensioni d'analisi, richiede il caricamento "ex novo" della base dati. Per questo, di volta in volta, occorre valutare il trade-off tra le due esigenze e adottare a seconda dei casi lo Star Schema o lo Snowflake Schema.
Va inoltre aggiunto che, per modelli molto complessi, con diverse fact table che condividono numerose dimensioni, diventa difficile adottare la tecnica dimensionale, anche perchè non è sempre chiaro quali siano le tabelle dei fatti e quali quelle delle dimensioni.

Per il disegno concettuale dell'Enterprise Data Warehouse appare preferibile, quindi, un modello dati Entity-Relationship opportunamente denormalizzato.
La modellazione dimensionale è invece senz'altro più idonea per la rappresentazione dei dati dei Data Mart.
I modelli dimensionali possono essere implementati sia su DBMS relazionali sia su specifici DBMS multidimensionali.

3.3.1 Strumenti di modellazione dati

Gli strumenti di modellazione dei dati sono prodotti che consentono di disegnare, utilizzando le più diffuse convenzioni rappresentative (ERA, Chen, ecc), i dati e definirne le caratteristiche, a diversi livelli di dettaglio: schemi concettuali, logici, fisici, garantendo anche la consistenza delle informazioni durante la trasformazione da un livello rappresentativo al successivo.
Inoltre tali strumenti possono prevedere la possibilità di recuperare gli schemi dati a partire da data base fisici.
Nell'ambito della modellazione dati ricadono le seguenti attività:

- Documentazione delle basi dati dei sistemi operazionali. Questa attività si rende necessaria per censire i dati a disposizione, esplicitarne la semantica e capirne le relazioni. Le basi dati operazionali possono essere state realizzate con DBMS diversi (relazionali, ma anche gerarchici o reticolari).
- Analisi dei dati di interesse e risoluzioni di eventuali conflitti. Questa attività consente di individuare i dati di interesse, strutturarli ed evidenziare eventuali conflitti che possono presentarsi se lo stesso dato esiste in DataBase distinti ovvero più volte sullo stesso DataBase.
- Progettazione delle basi dati. Questa attività prevede la definizione delle strutture delle basi dati dell'ambiente DataWarehouse
- Gestione delle basi dati. Questa attività consente di gestire la normale evoluzione delle strutture delle basi dati in oggetto.

Per effettuare le attività precedentemente indicate è indispensabile l'utilizzo di tool per :

- Controllare la complessità del progetto. Il numero di dati da trattare è elevato.
- Rappresentare con formalismi rigorosi (modelli[1]) la struttura dei DataBase, qualunque sia il DBMS su cui è realizzata.
- Rappresentare con formalismi rigorosi un modello concettuale dei dati, svincolato da aspetti di carattere tecnologico, che astragga ed espliciti le informazioni gestite a livello fisico.
- Mantenere la tracciatura tra i dati fisici ed i dati concettuali.
- Gestire la struttura della base dati.
- Produrre una documentazione flessibile ed integrata con l'ambiente di office automation.
- Mantenere aggiornata la documentazione a fronte di modifiche intervenute sulla base dati.

L'utilizzo di questi tool consente da un lato di dominare l'elevato quantitativo di informazioni da gestire mediante automatismi, meccanismi di astrazione e documentazione prodotta automaticamente, dall'altro di elevare a patrimonio aziendale conoscenze sui dati spesso detenute da fornitori o da singoli.

I tool presi in considerazione per effettuare le attività in oggetto sono stati :
- La suite COOL:BusinessTeam e COOL:DBA (ex Cayenne) della Sterling Software;[2]
- COOL:Biz della Sterling Software;
- ER-WIN della Computer Associates;
- Designer della Oracle.

che rappresentano i prodotti più diffusi sul mercato italiano.

Tutti i tool presentano un adeguata facilità di utilizzo ed una soddisfacente interfaccia ed integrazione con l'ambiente Windows.

Di seguito è presentata una matrice che illustra la presenza delle caratteristiche di interesse per le esigenze del progetto all'interno dei tool analizzati.

Tool/feature	BusinessTeam+DBA	BIZ	ER-WIN	Designer
Gestione di RDBMS	X[3]	X	X	X
Gestione di altre DataBase di interesse (DL1,	X			

[1] Per modello si intende un insieme di oggetti descritti in modo strutturato e organizati all'interno di diagrammi
[2] Con le dizioni COOL:Business Team e COOL:DBA si intende una serie di prodotti, che nel tempo, hanno subito varie denominazioni.
[3] X=copertura totale esigenza; O=copertura parziale esigenza;

flat file)				
Gesione Modello fisico	X	O	X	X
Gestione Modello concettuale	X	X		O
Integrazione modelli concettuali	X	X		X
Tracciabilità corrispondenza dati fisici e concettuali	X			O
Capacità di allineamento tra modello concettuale e fisico	X	O		O
Features per il controllo di correttezza/completeza modeli definiti	X		O	O
Personalizzazione della documentazione	X	X	X	X
Featutes specifiche di reverse engineering	X			
Lavoro multiutente		X	O[4]	X
Ambiente aperto[5]	X	X	X	

Il tool più adeguato per effettuare le attività illustrate nei paragrafi precedenti e la suite Business Team + DBA della *Sterling Software*. In particolare la suite di tool (ex Bachman, ex Cayenne) comprende:

- COOL:BusinessTeam (con annesso il modulo Report)
- COOL:DBA (con annesso il modulo Mapper)

I motivi salienti della scelta sono legati ai seguenti aspetti :

- I tool COOL:BusinessTeam e COOL:DBA sono gli unici che consentono la definizione di un reale modello astratto/concettuale dei dati, garantendo la tracciabilità tra dato concettuale e corrispettivi dati fisici (anche su DataBase distinti);
- I tool COOL:BusinessTeam e COOL:DBA sono gli unici che consentono di catturare informazioni da DBMS non relazionali;
- Il tool ER-Win non ha la possibilità di definire un modello concettuale dei dati e di conseguenza non supporta le attività di Data Administration;
- Il tool COOL:Biz non offre la possibilità di una completa gestione del ciclo di re-engineering, in particolare ha minori capacità nell'interfaccia verso i DataBase e nella tracciabilità della corrispondenza tra dato concettuale e fisico.
- Il tool Designer offre minori capacità espressive e di funzionalità per la definizione di modelli concettuali e per la tracciabilità tra dato concettuale e fisico. Inoltre Designer ha un non ottimale supporto al reverse engineering di basi dati non Oracle.

Di seguito viene riportata una breve descrizione dei prodotti scelti ed un elenco più dettagliato delle caratteristiche della suite che ne hanno determinato la scelta.

[4] è un modulo a se stante

[5] designer prevede che l'installazione si fatta su un cliente oracle (quindi è legato alla presenza del DBMS Oracle)

BusinessTeam è un tool per l'analisi integrata di dati e processi. In particolare per i dati consente di definire modelli concettuali rappresentati mediante diagrammi Entity-Rrelationship arricchiti con alcuni meccanismi di astrazione determinati dalla possibilità di rappresentare gerarchie di entità, relazioni complesse di vario tipo, relazioni M:M, classificazioni dei dati. I modelli possono essere costruiti ex-novo oppure in modo automatico catturando le informazioni da basi dati esistenti (anche di tipo non relazionale) e procedendo con delle attività di astrazione (reverse engineering).
Un sistema esperto assiste in tempo reale i progettisti per tutto ciò che concerne la completezza, la consistenza e la correttezza formale dell'analisi, guidandoli nelle decisioni e proponendo soluzioni alternative.
BusinessTeam è in ambiente Windows ed ha il supporto OLE container, che permette di integrarlo con tutte le applicazioni aderenti allo standard OLE. Tutte le esigenze di reporting sono soddisfatte dalla componente *BusinessTeam Reports*, che permette agli utenti di creare documentazione completa, consistente e standard per tutte le fasi di analisi e disegno.

DBA è un tool per la modellazione ed il disegno dei database. Il tool consente di catturare e generare basi dati per tutti gli ambienti relazionali in maniera completa e/o incrementale.
Un modulo addizionale (Mapper) permette di mantenere automaticamente allineati i modelli di business (definiti in BusinessTeam) e quelli delle basi dati (definiti con DBA).
COOL:DBA supporta direttamente, ed in maniera completa, il recupero, il disegno e la generazione degli ambienti Oracle, Sybase, e DB2 utilizzando schemi e costrutti che sono peculiari per ciascuno di questi tre DBMS. Questa soluzione è particolarmente adatta per tutti coloro che vogliano mantenere un unico modello dei dati senza precludersi la possibilità di realizzare/catturare da DBMS diversi o distribuiti su piattaforme diverse. Inoltre possono essere creati/catturati database da tutti gli altri RDBMS via ODBC.
COOL:DBA aiuta gli utilizzatori ad evitare gli errori ed ad assicurare la congruenza nei modelli consentendo in tal modo di aumentare la produttività e l'efficacia dei progetti. Tutto questo grazie alle regole di disegno implicite nel sistema e all'efficace " *Expert Advisor* ", un metodo on-line per il disegno e la realizzazione di modelli e basi dati.

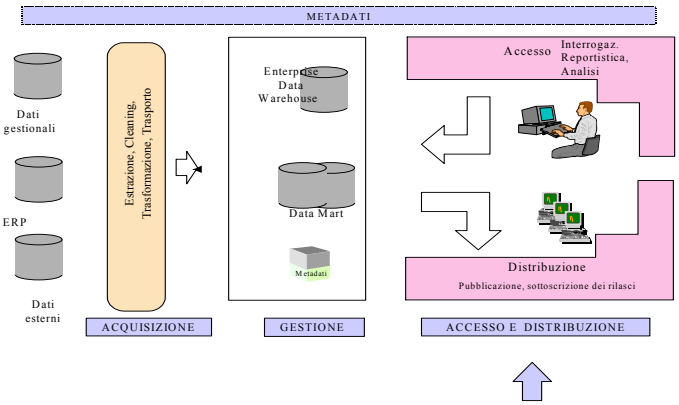
Alcune peculiarità che rendono la suite prescelta particolarmente adatta per le attività di Re-Engineering sono:

- Possibilità di interfacciare la gamma di DBMS di interesse, ovvero tutti i DBMS relazionali ed il DBMS gerarchico IMS/DL1.
- Possibilità di tenere separati gi aspetti implementativi da quelli di analisi. Infatti, è possibile definire un modello dipendente dall'ambiente tecnologico, contenente tutti gli oggetti del DBMS da documentare e le loro caratteristiche, ed un modello

indipendente da tale ambiente che riflette la visione dell'utente. In particolare il kit di tool consente la massima flessibilità nei processi di astrazione che conducono alla realizzazione del modello concettuale, ma allo stesso tempo conserva automaticamente la corrispondenza tra dato fisico e dato concettuale.

- Funzionalità di integrazione dei modelli concettuali mantenendo la tracciatura tra dati concettuali e fisici. Questa funzionalità consente di costruire i modelli dell'ambiente DataWarehouse integrando e semplificando quelli corrispondenti all'ambiente operazionale ottenuti con attività di reverse engineering operate su DBMS etrogenei.
- Funzionalità specifiche per il reverse engineering. E' possibile ricostruire automaticamente, a scopo documentativo, le relazioni esistenti all'interno di DBMS relazionali gestite via software e non definite nel DBMS (la mancanza di referential integrity nei database è in particolare molto diffusa in applicazioni DB2 datate).
- Funzionalità che consentono la verifica automatica della correttezza e completezza dei modelli ottenuti.
- Funzionalità che consentono di aumentare la produttività nella definizione dei modelli, automatizzando operazioni massive.
- Documentazione completamente personalizzabile in ambiente MS-Access e aggiornabile automaticamente a fronte di modifiche apportate sulle basi dati e/o sui modelli.
- Diffusione nel mercato italiano, e referenze su progetti complessi similari a quello in oggetto (INAIL, SOGEI, ISTAT, INA, TELECOM, TIM….).

3.4 *Accesso ai dati*



Le esigenze di accesso ai dati in un ambiente decisionale sono diversificate e, in linea con tali caratteristiche, il panorama di soluzioni possibili è ampio e notevolmente differenziato.

Tale diversificazione può essere riscontrata addirittura nell'ambito della stessa soluzione di Data Warehouse, in quanto legata alla diversità stessa delle attività degli utenti finali. Ad esempio è usuale che in una soluzione siano previste funzionalità per pochi utenti che svolgono attività di analisi "spinta" sui dati e funzionalità di reportistica più standard per molti altri utenti, la cui attività richiede la sola conoscenza delle informazioni.

In termini di architettura di accesso, la soluzione a tre livelli con tecnologia Web è da ritenersi preferibile, come descritto nel successivo paragrafo.

In assoluto è di sicuro auspicabile l'utilizzo di tool già disponibili sul mercato nel rispondere alle esigenze che nascono in tale area, per evidenti ragioni di economicità, innovazione tecnologica e integrazione delle soluzioni.

L'uso di tool non esclude l'esigenza di programmazione, nell'accezione di customizzazione del tool alle esigenze del progetto.

I prodotti disponibili possono essere raggruppati in quattro categorie:

- Strumenti di query e reporting
- Strumenti EIS
- Strumenti OLAP
- Strumenti di Data Mining

I paragrafi successivi sono indirizzati ad un approfondimento delle singole categorie.

3.4.1 Strumenti di query & reporting

Con tale dizione si intendono tool che consentono di effettuare interrogazioni sui dati anche da parte di utenti che non posseggono specifiche conoscenze di programmazione o di linguaggio SQL.
Molto semplici da usare, forniscono all'utente una rappresentazione grafica delle strutture dati e consentono la generazione e il riutilizzo di report.
In sintesi, forniscono un sottoinsieme delle funzionalità dei prodotti OLAP.

Rientrano in questa categoria, fra gli altri, i seguenti prodotti:

- Cognos Impromptu
- Platinum Forrest&Trees

3.4.2 Strumenti EIS

Gli strumenti EIS, **E**xecutive **I**nformation **S**ystem, sono prodotti sofisticati di presentazione dei dati rivolti principalmente al top management. Questi tool consentono di predisporre un'interfaccia grafica che rende possibile la navigazione sui dati attraverso menù personalizzati. L'interazione dell'utente è quindi limitata alle funzionalità definite in fase di analisi e realizzazione del progetto.
In genere le interrogazioni possibili operano su dati aggregati, proprio per la caratteristica di sinteticità dell'informazione da fornire.

Tool EIS disponibili sul mercato sono:

- Platinum Forrest&Trees
- SAS Institute Sa/EIS Server

3.4.3 Strumenti OLAP

L'acronimo OLAP**, O**n **L**ine **A**nalytical **P**rocessing, individua una classe di prodotti che consentono una "navigazione" nei dati (data surfing), fornendo strumenti di facile uso che mettono l'utente in grado di formulare in autonomia le interrogazioni.
Lo strato semantico fornito svincola l'utilizzatore dalla conoscenza delle basi dati, sia nei termini di localizzazione che di organizzazione.

Tipica di questa classe di prodotti di accesso è la visione multidimensionale dei dati, che consente la possibilità di analisi secondo percorsi (dimensioni) diversi (tempo, dislocazione geografica, etc.).
Il panorama degli strumenti OLAP è molto ampio; i produttori che operano nell'area sono circa un centinaio.
E' opportuna una precisazione sui concetti di ROLAP e MOLAP.

Con il termine ROLAP (Relational OLAP) si identificano strumenti che operano su dati memorizzati in database relazionali. In molti casi i dati sono memorizzati in strutture denormalizzate quali lo star schema e in forma aggregata.

I tool ROLAP si possono dividere in due categorie, a seconda che l'architettura di riferimento sia a due o tre livelli. Per entrambe le categorie si forniscono degli esempi:

ROLAP Client Tools (architettura a due livelli):
. Business Objects
. Oracle Discoverer

ROLAP Analytical Server (architettura a tre livelli):
. Informix Metacube
. Business Objects Web Intelligence
. Microstrategy DSS Server

Gli strumenti MOLAP (Multidimensional OLAP) si basano su un database proprietario, alimentato a partire dalle aree di staging o dall'Enterprise Data Warehouse.
Ad oggi non esiste uno standard di accesso a dati in forma multidimensionale così come esiste per i database relazionali (SQL). All'atto del caricamento, da effettuarsi con cadenza periodica, i dati vengono inseriti in strutture multidimensionali, con ottime garanzie di performance. Di contro dimensioni di analisi non previste al momento del caricamento non possono essere effettuate a meno di una successiva operazione di alimentazione.
Gli strumenti Molap sono realizzati secondo due diversi principi architetturali, che diremo Client e Server Molap.
Le due tipologie si differenziano nel fatto che nei Client Molap il database multidimensionale è residente sulla macchina client, nell'architettura Server Molap il database multidimensionale è ospite di un server.

Un esempio di prodotto Molap con architettura Client è Cognos Powerplay.

Esempi di prodotti Molap con architettura Server sono:
. IBM DB2 Server
. Hyperion Essbase Server
. Microsoft OLAP Server
. Oracle Express Server
. Seagate Holos

Punti di forza degli strumenti ROLAP sono l'interfaccia standard (SQL), l'accesso a qualsiasi database, la flessibilità di analisi, la scalabilità; viceversa i prodotti MOLAP hanno da sempre il loro punto di forza nelle prestazioni.
Negli ultimi anni si è assistito all'inserimento nei database relazionali di funzionalità specifiche per l'analisi decisionale, cosa che di fatto ha di molto ridotto il gap prestazionale fra le due tecnologie.
Molti produttori OLAP stanno variando la loro offerta fornendo soluzioni in cui cooperano prodotti MOLAP e ROLAP (soluzioni HOLAP, Hybrid OLAP).
Le soluzioni MOLAP sono proprietarie, meno flessibili, richiedono maggiore attività di avviamento e manutenzione rispetto alle soluzioni ROLAP.

Il loro uso non va escluso, anche se sicuramente deve essere valutato con attenzione.

Gli strumenti OLAP sono i più diffusi nell'eterogeneo panorama dei tool di accesso ai dati, in quanto forniscono funzionalità "spinte" di analisi dei dati, ma anche pregevoli funzionalità grafiche e di reporting.
Queste caratteristiche rendono tali prodotti d'interesse di una classe molto ampia d'utenza.
Normalmente l'utente analista predispone il report, consentendone poi l'utilizzo agli altri utenti che, a seconda dei casi, lo visualizzano semplicemente o ne richiedono magari un aggiornamento personalizzando certi parametri d'interesse.
L'utilizzo di tali tool rende in generale autosufficienti gli utenti nella capacità di localizzare, comprendere e accedere ai dati. Ciononostante, l'esperienza consiglia di prevedere periodi di affiancamento iniziali finalizzati a cogliere appieno le potenzialità offerte dalla soluzione e più in generale a consentire una evoluzione 'naturale' verso un modo nuovo di vedere e utilizzare i supporti informatici.
L'esplosione del fenomeno WEB ha fatto sì che di questi prodotti, nati in origine come soluzioni client/server, ne sia oggi disponibile anche una versione Web oriented.
La versione Web ha spesso funzionalità ancora limitate rispetto alla soluzione client/server.
La tendenza del mercato è quella di fornire prodotti con funzionalità equivalenti in architettura Client/Server e architettura Web.

3.4.4 Strumenti di Data Mining

Il Data Mining è un processo di analisi finalizzato alla identificazione di relazioni fra i dati ed alla costruzione di modelli previsionali di comportamento.
Questi tool, pertanto, sono indirizzati alla ricerca, a partire dai dati elementari, di regole, fenomeni, fattori che influenzano i risultati di un processo e a costruire modelli previsionali.

Rientrano in questa categoria, fra gli altri, i seguenti prodotti:

- Angoss KnowledgeSEEKER
- IBM Intelligent Miner
- PILOT Discovery Server
- SAS System

3.4.5 Tecnologia WEB per l'accesso al Data Warehouse

L'architettura prescelta per le soluzioni di Data Warehouse è quella a tre livelli.

Tale architettura, descritta dalla figura 2, prevede di implementare la logica applicativa su un servente (Application Server) distinto dal quello in cui vengono gestiti i dati (Data Server). Il Client svolge esclusivamente funzioni di presentation ed è pertanto definito "thin" (leggero).

Questa architettura si contrappone a quella Client/Server a 2 livelli che pone invece la logica applicativa sul Client e utilizza il servente semplicemente per accesso al database relazionale (fig.1).

Le configurazioni a tre livelli stanno rapidamente diventando lo standard dominante nel disegno delle applicazioni, trainate anche dalle potenzialità messe a disposizione dalla tecnologia WEB. Quest'ultima ha infatti dato slancio allo sviluppo di prodotti di middleware 'general purpose', semplificando l'implementazione e il partizionamento dell'applicazione.

Di conseguenza si assiste ad un'espansione del numero e della qualità dei tool di sviluppo e del middleware per le applicazioni a tre livelli. Lo stesso vale per i pacchetti applicativi ERP (*Baan, Sap e PeopleSoft*) che si stanno rapidamente convertendo all'architettura a tre livelli.

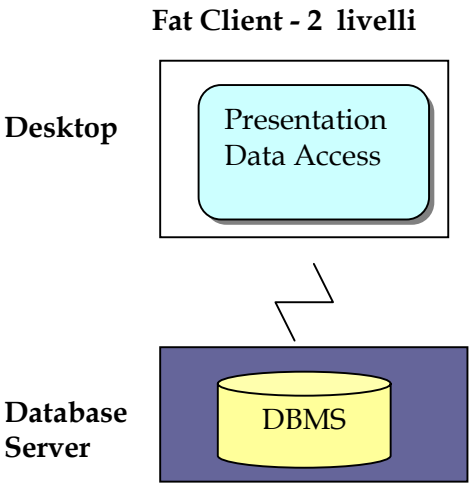


FIG.1

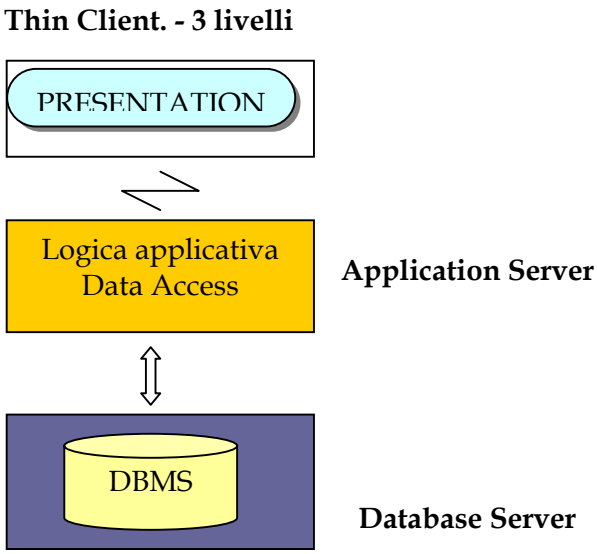


FIG.2

La tecnologia WEB traduce i tre livelli sopra descritti di Desktop, Application Server e Database Server rispettivamente in Web browser, Web Application Server e Data Base Server (in grado di gestire dati strutturati e non).

Il vantaggio dell'architettura a tre livelli è quello di consentire l'accesso da client leggero, configurato in modo standard, senza la complessità di gestire aspetti di aggiornamento e configurazione del SW di un cospicuo parco client.

Tutti i fornitori stanno caratterizzando la loro offerta con strumenti "Web enabling".

Il valore aggiunto dalla tecnologia WEB nel Data Warehouse consiste nella facilità di diffondere, distribuire e rendere disponibili informazioni aziendali (ad esempio, report, risultati di interrogazioni, analisi, documenti) ad utenti interni ed esterni all'Organizzazione. Inoltre offre, ad utenti collegati via *browser*, la possibilità di abbonarsi ad informazioni (*ordering*), di programmare il *delivery* di informazioni dipendenti da eventi e di pubblicare (*publishing*) informazioni.

La distribuzione e circolazione così capillare delle informazioni all'interno dell'Amministrazione è un volano fondamentale per l'innalzamento della produttività individuale e della qualità del lavoro.

Esempi di prodotti per Web publishing e delivery di report predefiniti sono:

- Business Objects Broadcast Agent
- Oracle Reports
- Seagate Crystal Reports Web Access
- Cognos Impromptu Web Reports
- Knosys ActiveBooks

Per l'esecuzione di analisi OLAP sui dati basate su WEB, tra gli altri possiamo citare

- Oracle Express Web Agent
- Business Objects ActiveX Report Reader
- Cognos Power Play
- Seagate Info7 Web Access
- Microstrategy DSS Web

Tra i tool OLAP basati su Java o ActiveX

- WEB Intelligence
- Infospace SpaceSQL and SpaceOLAP

Nel Middleware di mercato che supporta la tecnologia WEB, al fine di rendere fruibile l'informazione da parte dell'utente destinatario, diventano sempre più maturi gli strumenti di "information portal" (ad esempio VIT Delivery Manager).
Gli strumenti di information portal svolgono funzioni di publishing e delivery. Inoltre prevedono funzioni di gestione, che li rendono in grado di rappresentare un punto unico di ingresso e navigazione nella struttura informativa e di integrarsi con altri strumenti specializzati.

3.4.6 Considerazioni sull'accesso e la distribuzione

Analizzando l'eterogeneità delle esigenze, risulta da subito evidente che non esiste un unico tool in grado di rispondere a tutte le esigenze e di rappresentare la soluzione migliore in tutte le occasioni.

Ciononostante è sicuramente possibile identificare dei prodotti di riferimento per le varie categorie funzionali che rappresentano una risposta adeguata nella maggioranza dei casi.

Limitare il numero di prodotti utilizzati appare essere un obiettivo importante, per i vantaggi significativi che ne scaturiscono per l'utente finale e per l'Amministrazione nel suo complesso.

L'utente finale ne trae benefici in termini di semplificazione del posto di lavoro, modalità di accesso ai dati e processo formativo cui si deve sottoporre; tali benefici comportano per l'Amministrazione risparmi diretti e indiretti legati alle semplificazioni introdotte.

Attualmente nell'ambito del Ministero del Tesoro sono in uso i prodotti OLAP (Rolap) Business Objects e Oracle Discoverer.

Entrambi sono in possesso dei requisiti che valutiamo essenziali per un tool di data access:

- integrazione con i prodotti Data Warehouse esistenti presso l'Amministrazione
- soluzione web oriented o portabilità web
- integrazione con i prodotti di "office automation"

Inoltre va considerato il loro posizionamento di mercato.
Il tool Business Objects della omonima società ha il market share più alto a livello europeo; fa inoltre parte dell'offerta di Data Warehouse di IBM. Business Objects mette a disposizione un congruo numero di interfacce verso data base relazionali e prodotti MOLAP.
Oracle Discoverer è un tool con forte specializzazione per l'accesso a database Oracle.

I due tool sono adeguati sia per architetture a due livelli che a tre livelli.
Si reputa che i due rispondano a circa il 70-80% delle esigenze utente. Per casi particolari (ad esempio, il cruscotto aziendale-EIS) potranno essere valutati altri tool, così come non si esclude il ricorso a tool MOLAP.
Esigenze di query & reporting possono essere risolte facendo ricorso ai suddetti tool OLAP.

Per ultimo si riprendono due concetti attuali nell'ambito DWH: Business Intelligence e Information Delivery.
Col concetto di Business Intelligence si intende la nuova frontiera di queste tecnologie, volte a superare lo stato attuale di prodotti di accesso generalizzato a dati per innestarvi nuove funzionalità preconfezionate specializzate per l'area di applicazione, nonché integrazione con tool trasversali alle attività aziendali (ad esempio SAP e OFA- Oracle Financial Analyzer) .
Per Information Delivery si intende una caratteristica di talune soluzioni di prevedere la distribuzione automatica di report, prospetti, in generale informazioni verso gli utenti dell'organizzazione (push tecnology) tramite strumenti Web (channel broadcasting).

3.5 *La qualità dei dati*

L'iniziativa di Data Warehouse ha importanti conseguenze in merito al processo di miglioramento della qualità dei dati, sia di natura gestionale che direzionale. E' infatti chiaro che nel momento in cui le informazioni vengono consolidate nel data warehouse, è essenziale procedere ad una **verifica di qualità**, i cui esiti vanno concordati con le unità utente, al fine di prefigurare le azioni necessarie per un miglioramento del processo complessivo di acquisizione, elaborazione, diffusione del dato. In tal senso è urgente che l'Amministrazione si doti di una procedura tecnica ed organizzativa relativa al processo, da definire attraverso un manuale del controllo qualità dei dati, documento in cui tipicamente una organizzazione descrive come agisce per assicurare la qualità dei dati.
Si allega al presente documento il parere favorevole dell'AIPA (Adunanza del 27/01/2000) all'acquisizione del tool PowerCenter di estrazione, trasformazione e caricamento, in cui sono descritti i principali aspetti da trattare nel manuale della qualità dei dati.

Pur non avendo funzionalità specifiche legate alla "pulizia" dei dati, il tool Power Center consente di valutare il grado di qualità dei dati attraverso l'applicazione di condizioni di filtro e di migliorarne il livello attraverso processi iterativi di elaborazione.
Si intende verificare il grado di funzionalità specifica correlata alla pulizia dei dati del tool Power Center ed, in base ai risultati ottenuti, prevedere eventualmente l'acquisizione di uno specifico prodotto od eventuali servizi esterni per l'ottenimento dei risultati voluti in ordine all'attività di data cleaning.Al crescere delle informazioni memorizzate nel Data Warehouse e all'aumentare delle modalità con

3.5.1 Requisiti funzionali per i tool di qualità dei dati

I tool per la qualità dei dati debbono essere di ausilio alle attività di generazione e gestione delle basi informative durante tutto il loro ciclo di vita assicurandone la identificazione dei difetti di processo e degli errori sui dati.
Un tool di DQ (Data Quality) deve consentire l'attivazione di un processo che conduca alla misura, l'analisi e il miglioramento della Qualità dei Dati, ciò sarà ottenibile solo attraverso un monitoraggio nel tempo delle misure effettuabili sul contenuto reale degli archivi andando a verificare che i dati rispettino regole e caratteristiche prestabilite.

Le principali attività che dovranno essere consentite sono:

- Controllo (equivale all'analisi e alla individuazione delle anomalie e correzione, determina i livelli di qualità, in questa fase si decidono le politiche d'intervento);
- Monitoraggio (registra le variazioni dei livelli di qualità nel tempo a seguito degli interventi effettuati);
- Certificazione (descrive la correttezza del processo di qualità avviato);

permettendo interventi di migrazione, conversione e di re-engineering dei dati.

I requisiti funzionali che debbono caratterizzare i tool per la qualità dei dati riguardano:

- Identificazione di anomalie

  - Deve consentire una interfaccia ad un ampio insieme di tipologie di sorgenti di dati
  - Capacità di aprire in input più archivi contemporaneamente
  - Capacità di valutare la qualità del dato da una sorgente in input in termini di:
    - Frequenze di valori per livelli di campo distinti
    - Percentuali di valori corretti
    - Percentuali di valori che rientrano nel dominio di valori corretti
    - Percentuali di valori mancanti
    - Percentuali di valori non mancanti
    - Frequenze di formato (es: numero di occorrenze che un attributo assume in un particolare formato)
    - Numero di tipi di dati differenti per lo stesso campo (es: data/12/10/2000; 12 ott. 2000)
    - Frequenze di records con lo stesso identificatore;
  - Reportistica sugli errori rilevati durante i processi di accesso ai dati
  - Archiviazione dei record errati

  Deve essere possibile:
  - implementare regole di:
    - trasformazione,
    - consistenza,
    - correlazione
    - apparteneza
    -
  - verificare i vincoli di integrità definiti:
    - verifica che i valori dei campi equivalgano a valori predefiniti,
    - verifica di vincoli a livello di record,
    - verifica a livello di file o di insiemi di files

  - confrontare i dati Source con i dati Target:
    - confrontare a livello di record
    - confrontare a livello di campo

  - gestire l'errore e la relativa correzione
    - gestione multipla degli errori:
      - registrazione della informazione relativa all'errore in un file apposito
      - ignorare l'errore
      - segnalare l'errore con un "flag" per essere individuato.

      sviluppare reportistica sulle informazioni individuate nella gestione degli errori;
    - correggere l'errore in più di una modalità:
      - attraverso il sistema operazionale manualmente (off-line o on-line)
      - correzione automatica attraverso l'uso di un log
      - non correzione

- produrre report di controllo con semplicità
- produrre sintesi dei risultati delle verifiche effettuate
- evidenziare i records anomali (puntamento ai dati mancanti nel source o nel target)
- evidenziare i campi anomali (es: conversioni non valide)
- fornire la opportunità di estrazioni casuali ma rappresentativo (attraverso parametri definibili) degli archivi da analizzare
- deve consentire la manipolazione delle estrazioni casuali realizzate
- deve poter classificare i controlli di qualità eseguiti perché siano ripetibili nel tempo
- deve registrare le metriche e i processi di misura perché siano confrontabili nel tempo
- deve essere possibile clasterizzare metriche diverse per processi diversi di analisi da addattare a realtà distinte.

Il tool di data quality deve essere dotato di una interfaccia grafica che semplifichi la sua usabilità

- dovrà essere possibile l'uso di grafici che diano una informazione sintetica più efficace.
- dovrà essere possibile l'uso di grafici realizzati sulla base delle teorie più conosciute (es: Analisi di Pareto)

Prodotti per la qualità dei dati presenti sul mercato sono:

- Trillium Software – Brief (division of Harte-Hanks)
- Vality Technology – Integrity Data
- Group 1 Software Inc. - DQ Plus
- Informix Software - Quality Manager (formerly Prism Quality Manager),DataStage
- I.D. Centric(A Division of PostalSoft) - ACE
- Innovative Systems - Innovative
- QDB Solutions - QDB Analize, QDB Connect
- Ardent Software, Inc. - QM Quality Manager.

3.5.2 Considerazioni sulla qualità dei dati

I tool citati sono ancora poco utilizzati nel mercato italiano, sia per gli alti costi di acquisizione delle licenze, sia perché, sviluppati principalmente in America, implementano regole di codifica e standardizzazione non completamente aderenti alle nostre esigenze. Pertanto, pur proseguendo nell'attività di indagine e sperimentazione, si reputa conveniente realizzare i processi di qualità sui dati facendo ricorso allo sviluppo di software ad hoc o affidando in outsourcing a società specializzate i servizi di bonifica di specifici archivi.

3.6 *La sicurezza (logica)*

L'iniziativa di Data Warehouse, come ogni altra realizzazione informatica, non può prescindere dagli aspetti di sicurezza. Pertanto, anche la realizzazione del Data Warehouse sarà assoggettata alle policy di sicurezza in uso per gli ambienti esistenti del Ministero.

Nel seguito si tralascia la trattazione degli aspetti di sicurezza di tipo infrastrutturale, che peraltro non differiscono da quanto già in essere per gli ambienti esistenti, focalizzando l'attenzione sui dati.

Nell'analisi di sicurezza, è evidente che, per un sistema conoscitivo, non sia necessario procedere ad una analisi del rischio connesso alla criticità del dato gestito, in quanto il dato presente nel Data Warehouse ha, in modo intrinseco, le stesse caratteristiche del dato gestionale di cui esso è una copia. Pertanto dovranno essere aggiunte solamente le caratteristiche aggiuntive di cui deve essere corredato relative alle utenze che possono accedere al dato.

Tra i dati trattati nel sistema di Data Warehouse, ai fini della definizione delle caratteristiche di accessibilità, va fatta una suddivisione tra metadati e dati veri e propri.
I metadati, essendo solamente la documentazione e rappresentazione dei dati presenti in un sistema informativo, sono normalmente considerati di libero accesso all'interno di una organizzazione, cioè permeso a tutti gli utenti autorizzati alla visualizzazione di tali informazioni.
Infatti i metadati costituiscono il patrimonio comune di conoscenza e pertanto un valore aziendale che, pur con diversi livelli di percezione in funzione dei ruoli ricoperti dalle singole persone, può contribuire a facilitare la comunicazione permettendo di migliorare il lavoro degli utenti.

Diverso è invece l'approccio da seguire per i dati veri e propri, le cui caratteristiche di accessibilità sono solitamente diversificate in funzione della informazione che il dato rappresenta.

Ne consegue che, nella fase di analisi di ogni progetto che afferisce all'ambiente di data warehouse, nell'ambito delle specifiche di sicurezza, dovranno essere esplicitate anche le caratteristiche di accessibilità dei dati.

In particolare dovranno essere censiti:

proprietario del dato : è la funzione che istituzionalmente genera il dato gestionale sovraintendendo al processo di creazione, modifica, cancellazione del dato stesso. Il proprietario del dato fornisce le indicazioni sul tipo di accessibilità del dato stesso. Il proprietario è unico.

gestore del dato: è la funzione, che può essere diversa dal proprietario, che garantisce la qualità del dato gestionale, in termini di diponibilità, consistenza, aggiornamento.

tipo di accessibilità al dato: nel data warehouse i dati sono accessibili agli utenti in sola lettura. Per un dato si dovranno pertanto definire le caratteristiche di accesso degli utenti, che ad esempio possono essere:

- dato pubblico: tutti gli utenti che ne facciano richiesta (anche enti esterni)
- dato del Ministero: accesso consentito a tutti gli utenti del Ministero
- dato dipartimentale: accesso solamente agli utenti di un dipartimento (o di quali dipartimenti)
- dato di funzione (per funzione si intende Ispettorato, Ufficio , ecc.): accesso solamente agli utenti appartenenti ad una funzione (o di quali funzioni)
- dato nominativo: accesso solamente ad utenti nominativi

Tale classificazione sarà poi realizzata attraverso opportuni prodotti di gestione.
Pertanto, la presenza di un dato nel data warehouse non determina alcun automatismo nella disponibilita di un certo dato a nuovi utenti.

Si sottolinea comunque che la diffusione non dovuta delle informazioni a disposizione degli utenti esula dalle competenze e possibilità informatiche. Si vuole cioè evidenziare che non è possibile impedire, per via informatica, ad un utente, quando autorizzato a leggere dei dati, di copiare tali dati sulla propria stazione di lavoro, anlogamente a quanto avviene oggi per dati distribuiti tramite nastro, floppy, linea, supporto cartaceo.

La responsabilità del rispetto e dell'applicazione delle caratteristiche di accessibilità del dato all'interno del data warehouse è individuata nel responsabile Consip. Tale funzione, all'interno di Consip, sarà una attività non delegabile a fornitori esterni.

4. Considerazioni finali

Con questo documento si vuole indirizzare la necessità di affrontare in modo organico ed integrato i progetti di Data Warehouse.

La presenza delle diverse iniziative in essere nell'ambito dei Dipartimenti porta a considerare opportuno l'approccio "incrementale". Tale approccio consiste nella definizione di un Enterprise Data Warehouse come risultato di un percorso iterativo che prevede la realizzazione di Data Mart, secondo le priorità che verranno dettate dall'Amministrazione.

La integrazione dei diversi progetti è garantita dalla definizione di un modello informativo comune, che assicura la condivisione della semantica, la consistenza e la qualità dei dati.

Il modello informativo comune si arricchirà in modo iterativo dei modelli dati dei singoli progetti di data mart centralizzandoli in un unico repository. Tale modello potrà anche

essere consultato dagli utenti per condividere conoscenze sul patrimonio informativo del Ministero.

Nel caso in cui il fabbisogno informativo di un nuovo progetto di data mart, esaminato in dettaglio nella fase di analisi, sia già stato oggetto di acquisizione dai sistemi source da parte di altri progetti, se ne verificherà il riutilizzo direttamente a partire dall'Enterprise Data Warehouse o dalle aree di staging. Per la parte di fabbisogno non ancora disponibile, l'analisi porterà all'acquisizione del nuovo dato dai sistemi sorgente e rappresenterà un delta nel modello dell'Enterprise Data Warehouse.

Nel tempo, quindi, si costruirà una sorta di " dizionario dati"del Ministero, in cui saranno presenti le definizioni dei dati stessi (cosiddetti metadati).
Tra i metadati particolare attenzione andrà posta all'informazione circa la proprietà del dato e alle relative regole di trattamento e autorizzazione, in cui sono delineati gli aspetti relativi all'accessibilità dei dati.
Nella fase di definizione del significato semantico del dato e degli aspetti legati all'accessibilità è indispensabile il coinvolgimento del proprietario del dato.

Ogni singola iniziativa di data mart avrà quindi un collegamento al progetto di costruzione del Data Warehouse Integrato del Ministero, contribuendo a renderlo completo in modo iterativo.

Nello schema di riferimento proposto sono evidenziati i tre processi fondamentali di una architettura Data Warehouse:

- acquisizione
- gestione
- accesso e distribuzione

Dal punto di vista dell'infrastruttura tecnologica si predisporrà un ambiente centralizzato dedicato che dovrà offrire adeguate garanzie in termini di protezione, sicurezza e disponibilità dei dati e dei sistemi, basato su una macchina a tecnologia parallela (tra quelle indicate nel paragrafo relativo alle piattaforme HW).

I processi di acquisizione saranno localizzati esclusivamente su questo ambiente. Dovranno inoltre essere disponibili processi di gestione ed evntualmente distribuzione e accesso.

Quanto al SW, su tale sistema saranno installati il RDBMS (Oracle), il tool di estrazione, cleaning, caricamento e replica (Power Center della Società Informatica) e gli eventuali tool di accesso e distribuzione.

Il sistema rappresenterà il Data Server per l'Enterprise Data Warehouse e per i Data Mart dei progetti che abbiano un significativo volume di dati ed un utenza ampia come potrebbe essere nel caso della Banca Dati Progetti di Investimento Pubblico.

L'ambiente dedicato ospiterà, inoltre, le aree di staging comuni, cioè aree tecniche, non accedibili direttamente dall'utente finale. Su queste aree verranno consolidati gli output dei

processi di estrazione dai vari sistemi sorgente e centralizzate le operazioni di cleaning e trasformazione preliminari al popolamento dei Data Mart .

Per i Data Mart più contenuti è ipotizzabile l'utilizzo di Data Server dipartimentali su piattaforma Unix o Windows/NT.

Sui Data Mart risultano centrati i processi di accesso e distribuzione, che rendono il dato disponibile all'utente finale. L'architettura di riferimento proposta per tali processi è quella a tre livelli, con apertura alle potenzialità offerte dalla tecnologia WEB, come espresso nei requisiti di quasi tutti i Progetti del Ministero.
Analizzando l'eterogeneità delle esigenze dell'utente finale, risulta opportuno non restringere la scelta ad un unico tool di accesso e distribuzione. D'altra parte, limitare il numero di prodotti utilizzati è importante per i vantaggi significativi che ne scaturiscono per l'utente finale (modalità standard di accesso alle informazioni, conoscenza dello strumento) e per l'Amministrazione nel suo complesso.

Quindi, a meno di esigenze molto specifiche da valutarsi caso per caso, sono stati identificati due prodotti di riferimento: Business Objects (Business Objects) e Oracle Discoverer (Oracle). Essi, per la diffusione all'interno del Ministero e delle loro attuali caratteristiche tecniche, rappresentano una risposta adeguata nella maggioranza dei casi, come dimostra peraltro il loro posizionamento sul mercato.