Università degli Studi di Verona

Sistemi informativi

Riassunto dei principali argomenti

Davide Bianchi

May 16, 2019

Contents

1 Teoria dell'organizzazione

1.1 Introduzione

Iniziamo con alcune definizioni *estremamente* tediose.

Definizione 1.1.1 (Sistema informativo) *Il Sistema Informativo (SI) è la componente (sottosistema) di una organizzazione che gestisce le informazioni di interesse.*

Definizione 1.1.2 (Organizzazione) *Un'organizzazione è :*

- *il processo attraverso il quale tale insieme di persone viene strutturato secondo i principi di divisione del lavoro e coordinamento;*
- *il risultato del processo di divisione del lavoro e coordinamento.*

Definizione 1.1.3 (Azienda) *Un'azienda, nell'economia aziendale, è un'organizzazione di uomini e mezzi finalizzata alla soddisfazione di bisogni umani attraverso la produzione, la distribuzione o il consumo di beni economici.*

Definizione 1.1.4 (Organizzazione aziendale) *Il processo attraverso il quale l'insieme di persone che partecipano direttamente allo svolgimento dell'attività dell'azienda viene strutturato secondo i principi di divisione del lavoro e coordinamento.*

L'organizzazione aziendale ha sempre almeno i macro processi operativo e gestionale, e dispone di risorse materiali, umane e informative.

Definizione 1.1.5 (Tecnologie informatiche) *Insieme di sistemi, strumenti e tecniche predisposti per automatizzare il trattamento delle informazioni.*

Un sistema informativo aziendale è una collezione di elementi interconnessi che gesticono la raccolta, l'elaborazione e la restituzione di informazioni.

Un sistema produttivo aziendale è basato su *obiettivi* (output atteso), *input* ed *output* effettivi. Definiamo inoltre i concetti di *efficienza*, ovvero il costo di raggiungimento degli obiettivi, e di *efficacia*, ovvero il grado di raggiungimento degli obiettivi. A grandi linee:

$$\textit{Efficienza} = \frac{Output}{Input} \qquad \textit{Efficacia} = \frac{Output}{Obiettivi}$$

Efficienza ed efficacia subiscono impatti differenti rispetto ad una innovazione delle risorse tecnologiche. Nel caso dell'efficienza:

- Riduzione dei costi unitari;
- Aumento di produzione a parità di risorse;
- Incentivo alla crescita delle dimensioni organizzative;
- Maggiore complessità strutturale;
- Cambiamenti della struttura organizzativa.

Nel caso dell'efficacia:

- Più efficiente uso dei fattori produttivi a parità di volumi di produzione (economie di scopo);
- Razionalizzazione dell'uso di risorse;
- Maggiore efficienza spesso legata a differenziazione dei prodotti e all'ampliamento della gamma.

L'organizzazione è un sistema aperto, influenzato da variabili ambientali, che vengono riassunte nell'*incertezza ambientale*. L'incertezza ambientale determina i requisiti di capacità elaborativa della organizzazioni e l'adeguatezza del sistema informativo.

Definizione 1.1.6 (Capacità elaborativa) *Adeguatezza di un'organizzazione rispetto alle necessità di elaborare informazioni a essa imposte dai propri obiettivi e dal contesto in cui opera.*

L'ambiente è riassunto nel modello della piramide di Anthony, ovvero una piramide divisa in livelli gerarchici. Il layer più alto si occupa delle decisioni strategiche, quello centrale delle decisioni direzionali e quello basso di quelle operative.

1.2 Informazione come risorsa organizzativa

Caratteri principali. L'informazione è la vera risorsa nelle attività organizzative, infatti viene scambiata ed elaborata; è immateriale, non è facilmente divisibile, può essere soggetta ad obsolescenza e si autorigenera.

La capacità autorigenerativa dell'informazione permette di instaurare circoli virtuosi di generatori di conoscenza e di arricchimento delle informazioni disponibili, che si traducono in un incremento dei processi produttivi.

Overload e underload informativo. L'*overload informativo* è un aumento incontrollato dell'informazione disponibile, che eccede la capacità di elaborazione individuale, con conseguente rallentamento nell'elaborazione. L'*underload informativo* è invece una disponibilità di informazione al di sotto delle capacità individuali, con conseguente presa di decisioni in tempi brevi.

1.3 Sistemi informativi verticali e orizzontali

I sistemi informativi possono essere immaginati a due versi.

I sistemi informativi verticali sono stati i primi ad essere supportati dai sistemi informatici, tuttavia al crescere dell'incertezza i vertici sono sovraccaricati dai compiti decisionali.

I sistemi informativi orizzontali invece sono costruiti sulla delega delle decisioni e sui collegamenti sullo stesso layer che aumentano la capacità elaborativa (team di lavoro, task force).

2 Classificazione dei sistemi informativi

Vi sono varie possibili classificazioni dei sistemi informativi, ovvero:

- tipologie di SI disposti lungo la piramide aziendale (definizioni e funzioni attribuite a seconda del loro livello nella piramide);
- tipologie di SI diposti nelle varie aree gestionali dell'impresa.

2.1 SI disposti lungo la piramide aziendale

I SI disposti lungo la piramide di Anthony sono i seguenti:

1. *Transaction Processing Systems (TPS)*: gestione delle transazioni, quali ordini ecc. Sono alla base della piramide;
2. *Management Information Systems (MIS)*: sono al livello immediatamente sopra ai TPS e rappresentano periodicamente le informazioni raccolte dai TPS. Sono alla base del sistema di reportistica delle aziende;
3. *Decision Support Systems (DSS)*: affiancano il management delle decisioni di non routine e permettono di simulare ipotesi per verificare la validità di una gestione.
4. *Executive Information Systems (EIS)*: sono al vertice della gerarchia, aiutano i senior manager alla gestione.

2.2 Portafoglio applicativo

Il portafoglio applicativo è l'insieme delle applicazioni utili in azienda. È diviso in 3 segmenti principali:

- Portafoglio direzionale: insieme delle applicazioni informatiche a supporto dei cicli di pianificazione strategica;
- Portafoglio istituzionale: applicazioni informatiche per i processi di supporto all'amministrazione;
- Portafoglio operativo: applicazioni informatiche per i processi primari dell'azienda.

Il portafoglio istituzionale è un'area con elevate potenzialità di informatizzazione, a causa dei grandi volumi di dati che li coinvolgono e la forte proceduralità. Come conseguenza si hanno riduzioni nei tempi e nei costi di elaborazione, inoltre la pianificazione risulta più efficace.

Il portafoglio operativo contiene le applicazioni informatiche necessarie ai procedimenti coinvolti nella catena del valore di Porter, ossia:

Gestione materie prime →Trasformazione →Vendita →Distribuzione →Postvendita

ovvero la catena di azioni finalizzate a produrre valore per il cliente. Il portafoglio applicativo ovviamente è specifico di ogni settore industriale, e comporta un aumento della complessità gestibile nei processi aziendali, permettendo inoltre la sincronizzazione dei dati in un'azienda (basi di dati condivise).

Il portafoglio applicativo è andato informatizzandosi col passare degli anni, a cominciare dalle procedure per automatizzare attività singole, ai pacchetti MRP (Manufacturing Resource Planning), che contenevano i primi database e pacchetti integrati, ai CIM (Computer Integrated Manifacturing), che automatizzano interi segmenti produttivi, ai sistemi ERP (Enterprise Resource Planning) che consentono di gestire ogni fase produttiva e sfruttano architetture client-server e pacchetti integrati con un unico modello dati.

Negli ultimi anni sono andati sviluppandosi i sistemi CRM (Customer Relationship Management), che forniscono interi cicli di assistenza al cliente, gestioni avanzate di distribuzione, vendita e postvendita. Negli anni 2000 si è sviluppato l'E-Procurement, ovvero l'informatizzazione del buy-side delle imprese, e utilizza pacchetti per l'intero ciclo di acquisto e architetture basate su tecnologie web.

3 Ingegneria Sociale

Definizione di Ingegneria Sociale a caldo:

Definizione 3.0.1 (Ingegneria Sociale) *Manipolazione della naturale tendenza alla fiducia dell'essere umano, architettata e realizzata dall'ingegnere sociale con l'obiettivo ottenere informazioni che permettano libero accesso e informazioni di valore del sistema.*

La figura dell'ingegnere sociale mira a stabilire confidenza con la vittima, sviluppando ogni possibile scenario di difficoltà e preparandosi ad evaderlo. Prima di tutto ciò viene la fase di *footprinting*, ovvero di raccolta delle informazioni, l'analisi dell'azienda, dei suoi sistemi di comunicazione, della posta, ecc. I primi ingegneri sociali furono i *phreaker*, che utilizzavano la rete telefonica sfruttando i sistemi e i dipendenti dell'azienda per arrivare a dati sensibili.

La falla da sfruttare è quindi data da operatori umani, che spesso gestiscono le informazioni sensibili ma ignorano le procedure di sicurezza, magari non sono nemmeno consapevoli delle informazioni che stanno gestendo e che dovrebbero custodire. Ovviamente le vittime perfette per un attacco di ingegneria soociale sono le persone che non hanno nulla da perdere nel fornire informazioni sensibili, che sottostimano il valore delle informazioni, sottostimano le procedure di sicurezza oppure che non valutano le conseguenze delle proprie azioni.

3.1 Attacco tipico di Ingegneria Sociale

Fasi di un attacco di SE. Un generico attacco di SE si svolge nel seguente modo:

- una fase fisica di raccolta di informazioni attraverso persone, documenti e luoghi;
- una fase psicologica di impersonificazione e persuasione del personale adatto ad essere una tipica vittima

Fase fisica. Gli strumenti essenziali alla fase fisica sono gli strumenti di comunicazione più disparati. L'obiettivo di questa fase sono password, server e router, e si possono raggiungere tramite una giusta combo di uso della tecnologia (phishing, lancio di malware, ...) e interazione col personale (truffe telefoniche, dumpster diving, rovistare negli hdd dismessi...).

Fase psicologica. È necessario fare leva sulla fiducia che una persona è per inclinazione disposta a concedere, facendo leva sui bisogni primari dell'uomo (fisiologici, di sicurezza, ecc.), secondo la gerarchia di Maslow.

Gli attacchi di social engeneering sfruttano quindi le debolezze della persona singola, ossia la disponibilità e la fede che una persona è disposta ad affidare ad un possibile attaccante.

Definizione 3.1.1 (Phishing) *Tecnica di ingegneria sociale basata sul principio della supposta autorità che utilizza un messaggio di posta elettronica per acquisire informazioni personali riservate (password, dati finanziari, numero di carta di credito) con la finalità del furto di identità.*

Il phishing è basato sul concetto di *mail spoofing*, ossia sull'inviare mail a nomi di terzi, per il semplice motivo che la persona che manda la mail non è autenticata dal server di posta elettronica.

4 Ingegneria dei processi gestionali

I processi rappresentano il modo di operare in un'azienda. Dal momento che le tecnologie informatiche modificano il modo di operare in un'azienda, è necessario il processo di **Business Process Reengineering** (BPR), che mette in correlazione l'innovazione dei processi e dell'organizzazione aziendale tramite l'uso di strumenti informatici.

I processi possono essere materiali, informativi, oppure **Business Process**, ovvero un insieme di attività finalizzato alla realizzazione dell'interesse dell'azienda. In generale un processo aziendale è formato da attività, che, partendo da input definiti, producono l'output richiesto dal cliente. I processi sono flussi di attività che concatenano marketing, produzione e approvigionamento.

Un business process è un processo costituito da 4 elementi:

- attività;
- input (materiali o risorse di partenza);
- output (oggetti, materiali, servizi in uscita);
- clienti.

Considerando come è definito un processo, un'azienda è facilmente costituibile come un processo (catena del valore di Porter). La catena del valore di Porter è suddivisibile in 3 macro-strategie: una *buy-side* (acquisizione delle risorse), *in-side* (trasformazione delle risorse) e *sell-side* (vendita, distribuzione, postvendita).

Strategia buy-side. Include il rapporto con i fornitori, con una potenziale riduzione dei costi del materiale stesso. La strategia buy-side si appoggia ai sistemi di e-procurement, infrastrutture internet, mercati elettronici.

Strategia in-side. È mirata alla trasformazione dei processi interni dell'azienda. Questa strategia può arrivare ridurre il costo di funzionamento dei processi e migliorare il rapporto con il cliente. È appoggiata principalmente ai sistemi ERP.

Strategia sell-side. È orientata ai processi di marketing, vendita e distribuzione dei prodotti. La trasformazione dei processi si appoggia ai sistemi CRM e comporta un maggiore valore del prodotto percepito dal cliente e un abbattimento dei costi di produzione.

4.1 Classificazione dei processi

I processi sono classificabili nelle seguenti categorie:

- *intersettoriali:* gestiscono le pratiche di molteplici settori;
- *settoriali:* distinguono i vari settori;
- *aziendali:* processi di una specifica azienda o di una sua parte;
- *normativi e best-practice:* sono processi di riferimento e di giuda su come gli altri processi dovrebbero essere nelle aziende del settore.

Ogni processo è scomposto (dal macro al micro) nelle seguenti fasi:

- *Macroprocesso*: constituisce il primo livello di segmentazione dell'azienda, la catena del valore di Porter ne è un'esempio;
- *Processo*: illustrano il modo di operare dell'azienda;
- *Fase*: illustrano il modo in cui il processo è implementato (una fase è una tappa del processo);
- *Attività*: livello minimo di analisi normalmente adottato nello studio dei processi, sono operazioni fatte da singoli o pochi;
- *Operazione*: passi elementari necessari per eseguire una data attività (mai usate).

4.2 Griglia metodologica

La griglia metodologica è uno strumento di supporto alla progettazione dei processi. Questa metodologia comprende

- Descrizione delle variabili di analisi;
- Descrizione delle fasi di analisi;
- Identificazione degli strumenti di supporto alle analisi.

4.2.1 Variabili organizzative

La trasformazione dei processi per avere successo deve ruotare intorno ai perni dell'innovazione tecnologiche e organizzative. Le variabili organizzative sono i punti su cui bisogna lavorare per ottenere un successo (almeno teoricamente). Le variabili sono le seguenti:

- Flusso delle attività: sequenza di attività attraverso le quali il processo è svolto;
- Organizzazione del processo: è la divisione operativa del processo e come i singoli compiti sono mappati sui ruoli ;
- Competenza delle risorse umane che operano nel processo.

Flussi di attività. Il flusso delle attività determina la durata del processo, il livello di servizio e la qualità del prodotto. La modellazione dei flussi può essere ricondotta a diversi schemi: schemi di sequenza (che raccolgono solamente i caratteri delle attività da svolgere) oppure altri flussi più ricchi (che raccolgono altre variabili ma sono più complessi da realizzare).

Organizzazione. L'organizzazione è una variabile fondamentale nella fase di descrizione dei processi. Per descrivere un'organizzazione si possono utilizzare organigrammi, tabelle di proprietà e Linear Responsibility Charting (LRC). Le tabelle di proprietà sono delle tavole che elencano, per ogni organo aziendale, i compiti devoluti a tale organo, i processi svolti, gli organici ecc.

I Linear Responsibility Charting offre una visione tabellare delle reponsabilità organizzative. Lo scopo è, per ogni processo, identificare il ruolo svolto da ogni struttura aziendale (ruolo che può essere decisionale, esecutivo, di supporto...).

Risorse umane. Le risorse umane determinano la differenza tra il risultato effettivo di un processo e il massimo risultato teoricamente possibile in una data configurazione. Ovviamente l'innovazione tecnologica porta alla necessità di avere un insieme di figure altamente specializzate, acquisite dal mercato oppure anche riformando le figure già esistenti.

Analisi delle prestazioni. Ogni processo è valutato con un sistema di analisi delle prestazioni, comprendente:

- Pianificazione e controllo che fissa gli obiettivi di efficacia ed efficienza del metodo;
- Incentivazione e promozione che fissa gli obiettivi del processo e valuta il lavoro del singolo;
- Sistema dei valori che descrive gli obiettivi generali dell'azienda e decide i valori da incentivare nel lavoro (soddisfazione del cliente, produttività, ecc.).

4.3 Analisi dei processi

La metodologia di analisi dei processi è divisa in molteplici fasi. Gli approcci sono 2:

- *bottom-up*: ridisegno dei processi basato sul confronto con altre aziende;
- *top-down*: dati dei criteri di ottimizzazione, si lavora al disegno del processo seguendo tali criteri.

Alla base di entrambi gli approcci è necessario analizzare la situazione esistente, attraversando una fase preliminare di analisi della situazione di partenza, attraverso i passaggi di:

1. Identificazione dei processi (input, output, tipo, clienti del processo);
2. Dettagli del processo (diagrammi gerarchici, diagrammi di flusso, schede che descrivono le proprietà dei processi)
3. Incrocio processi/Unità organizzative (rilevazione delle strutture e dei ruoli, mappatura delle attività del processo);
4. Valutazione del processo (definizione dei parametri di funzionamento, giudizio sul valore dei prodotti, sia da parte del cliente che degli esecutori).

La fase successiva dell'analisi di un processo è composta dalle seguenti fasi:

- Confronto quantitativo e parametrizzazione (confronto tra aziende concorrenti);
- Confronto qualitativo (analisi delle divesità rispetto ai valori di mercato).

La terza ed ultima fase dell'analisi di un processo è data dalla **ridefinizione** del processo, ossia della *vision* che da una rappresentazione degli elementi fondamentali della soluzione proposta, solitamente basata su degli schemi best-practice.

La trasformazione dei processi, dovuta in larga parte all'applicazione di tecnologie informatiche, ha come effetto primario l'integrazione inter-funzionale o inter-organizzativa dei processi, in quanto aumenta al disponibilità di informazioni e supporta l'esecuzione di compiti individuali di natura decisionale.

5 Data warehouse

5.1 Introduzione

Iniziamo dando la definizione di *business intelligence*.

Definizione 5.1.1 (Business Intelligence) *Disciplina che consente a chi deve decidere in azienda di capire, attraverso soluzioni software, i fattori chiave del business e conseguentemente di prendere le migliori decisioni in quel momento.*

Sostanzialmente il ruolo chiave della business intelligence è quello di trasformare i dati aziendali in informazioni fruibili in maniera semplice, con un livello di dettaglio variabile.

Il passaggio da grandi moli di dati ad informazioni importanti e di un certo rilievo è operato dall'informatica, dal momento che negli ultimi anni le moli di dati sono diventate davvero enormi. A partire dagli anni 80 infatti iniziano a nascere i *decision support systems* (DSS), ovvero un insieme di tecniche e strumenti che consentono di estrapolare informazioni a partire da dati grezzi.

Qui nasce il data warehouse, ovvero un raccoglitore di informazioni che riorganizza i dati provenienti dalle sorgenti più disparate e li rende disponibili per analisi e valutazioni finalizzate alla pianificazione del processo decisionale.

Definizione 5.1.2 (Data Warehouse) *Una collezione di metodi, tecnologie e strumenti di ausilio al knowledge worker per condurre analisi dei dati finalizzate all'attuazione di processi decisionali e al miglioramento del patrimonio informativo.*

Le interrogazione al data warehouse sono di due tipologie:

- *OLAP (OnLine Analitycal Processing)*: interrogazioni di tipo analitico, leggono grandi quantità di record e calcolano dati di sintesi;
- *OLTP (OnLine Transactional Processing)*: interrogazioni di tipo transazionale, dove vengono letti e modificati piccoli gruppi di record.

Il processo di data warehousing presenta alcune caratteristiche fondamentali, quali:

- accessibilità ad utenti sprovvisti di particolari conoscenze informatiche;
- integrazione dei dati sulla base di un modello standard dell'impresa;
- flessibilità di interrogazione;
- sintesi per permettere analisi quanto più possibili efficaci;
- rappresentazione quanto più intuitiva possibile;
- completezza e correttezza dei dati manipolati.

5.2 Architetture di supporto

Le architetture di supporto al DW devono soddisfare alcuni requisiti fondamentali:

- *separazione*: tra elaborazione analitica e transazionale;
- *scalabilità*: le dimensioni e le performance dell'architettura devono poter scalare bene con l'aumento della dimensione della mole di dati;
- *estendibilità*: deve essere possibile estendere il sistema con nuove tecnologie senza doverlo riprogettare del tutto;
- *sicurezza*: relativamente al controllo degli accessi;
- *amministrabilità*: la complessità dell'attività di amministrazione non deve risultare eccessiva.

Architetture a un livello. L'unico intermezzo tra gli strumenti di utility (reportistica, OLAP) è il middleware.

Architetture a 2 livelli. Le architetture a 2 livelli sono strutturate in maniera differente, in quanto ci sono anche i livelli di alimentazione e di warehouse. La particolarità di questa architettura è data dai *data mart*, ovvero un sottoinsieme del data warehouse principale, relativi a sezioni particolari dell'azienda. I data mart sono utili come blocchi costruttivi del DW principale, inoltre, essendo di dimensioni minori, consentono migliori prestazioni.

In alcuni casi si preferisce un'architettura con i data mart indipendenti dal DW primario, il che semplifica la progettazione ma rende lo schema più complesso a livello di accesso ai dati, e può determinare inconsistenze tra data mart.

Le architetture a 2 livelli presentano sostanziali vantaggi, quali:

- a livello del DW sono sempre disponibili informazioni, anche se alle sorgenti non lo sono;
- l'interrogazione analitica sul DW non interferisce con quelle transazionali sul database operazionale;
- l'organizzazione logica del DW è multidimensionale, non relazionale o semistrutturata;
- a livello del DW sono ottimizzabili le interrogazioni ed in generale le prestazioni.

Architetture a 3 livelli. Hanno in più i dati *riconciliati*, che sono ottenuti dopo una serie di controlli di consistenza e pulizia. Il vantaggio dei dati riconciliati è che forniscono un modello unico, eliminando le problematiche per l'azienda di estrazione dei dati dalle sorgenti. D'altro canto, si introduce un elemento di ridondanza con i dati della sorgente.

5.3 ETL (Extraction, Transformation, Loading)

Il ruolo degli strumenti ETL è quello di alimentare una sorgente dati di buona qualità che possa a sua volta alimentare il DW (*riconciliazione*). La riconciliazione avviene in due momenti distinti, ovvero quando il DW viene popolato per la prima volta e quando viene aggiornato periodicamente.

Gli strumenti ETL operano attraverso 4 fasi:

1. **estrazione**: l'estrazione *statica* viene fatta quando il DW viene popolato per la prima volta, quella *incrementale* viene fatta dopo ogni aggiornamento, basandosi su cosa è cambiato nel frattempo;
2. **pulitura**: si migliora la qualità dei dati da elaborare, rimuovendo dati mancanti, duplicati e incosistenze;
3. **trasformazione**: conversione dei dati dal formato operazionale a quello del DW, i dati vengono normalizzati, convertiti nei formati supportati dal DW, e vengono sintetizzati;
4. **caricamento**: i dati vengono importati nel DW con il *refresh* (riscrittura totale dei dati) o con l' *update*, quando si devono fare aggiornamenti controllati e di dimensioni ridotte.

5.4 Il modello multidimensionale

È il modello fondamentale per la rappresentazione e l'interrogazione ai dati. È strutturato su un cubo i cui lati sono una dimensione dei dati da analizzare, ogni cella contiene la misura numerica da analizzare.

5.5 Tecniche di analisi dei dati

Una volta che i dati sono stati trasformati e ripuliti, occorre scoprire come trarne il massimo vantaggio informativo. Esistono 3 differenti approcci:

- data mining;
- reportistica;
- OLAP.

OLAP. È il principale metodo di analisi dei dati in un modello multidimensionale. È estremamente dettagliato grazie alla specifica di un percorso di navigazione, ossia un path di operatori che prendono in input il risultato dell'interrogazione precedente. Ogni risultato di interrogazione è ancora un modello multidimensionale. Gli operatori sono:

- roll up;
- drill down;
- slice and dice;
- pivoting;
- drill across.

Data Mining. È un'attività orientata alla scoperta di dati nascosti, in particolare quando la mole di dati è molto grande. Il data mining raccoglie tecniche di machine learning e AI al fine di ricercare caratteri particolari a partire dalla mole di dati, quali:

- ricerche di mercato;
- efficacia del marketing;
- pianificazione aziendale e di investimenti;
- riconoscimento di attività fraudolente.

Un elemento fondamentale del data mining è dato dalle **regole associative**, ossia un insieme di regole che consentono di stabilire delle relazioni di implicazione all'interno della base di dati (gruppi di affinità). In tal modo è possibile costruire pubblicità mirate, ad esempio. Altri caratteri fondamentali del data mining sono:

- *clustering:* raggruppamento di oggetti per carattere comune in un ridotto numero di insiemi;
- *alberi decisionali:* alberi di decisione per la classificazione dei rapporti di causa effetto di un dato evento;
- *serie temporali:* individuazione di pattern ricorrenti in sequenze di dati complesse (rilevazione di anomalie, analisi di percorsi web, ecc...).

6 Ciclo di vita del Data Warehouse

Generazione del DW. Nella generazione del DW si distinguono 2 approcci:

- *top-down*: basato sull'analisi dei bisogni dell'azienda, per poi realizzare il DW nella sua interezza, ha costi onerosi e lunghi tempi di realizzazione, ma garantisce ottimi risultati;
- *bottom-up*: la costruzione avviene in maniera incrementale, ha costi ridotti e tempi di realizzazione brevi, ma determina una visione parziale del dominio di interesse.

Qualunque sia l'approccio scelto, si segue sempre il path del tipo:

1. Pianificazione;
2. Progettazione dell'infrastruttura;
3. Progettazione e sviluppo del data mart.

Analisi e riconciliazione delle sorgenti. L'analisi e la riconciliazione delle sorgenti attraversa 2 fasi, la *ricognizione* e la *normalizzazione*.

Vi è prima la fase di *integrazione*, ovvero l'analisi delle sorgenti e il mapping, ovvero la correlazione di concetti delle sorgenti a schemi del nostro DW. La fase di integrazione è divisa in 4 step, ossia:

1. *Preintegrazione:* definizione della strategia di integrazione;
2. *Comparazione degli schemi*: analisi degli schemi iniziali per identificare conflitti e correlazioni tra concetti;
3. *Allineamento degli schemi:* risoluzione dei conflitti rilevati precedentemente;
4. *Fusione degli schemi:* si fondono gli schemi ottenuti con lo scopo di formare un unico schema riconciliato.

Analisi dei requisiti. Lo scopo di tale analisi è quello di raccogliere le esigenze di utilizzo del data mart. L'analisi dei requisiti è un processo fondamentale in quanto influenze le decisioni riguardanti la progettazione del database, l'architettura del sistema, i piani di avviamento e manutenzione.

6.1 Progettazione concettuale

Al contrario dei database relazionali, per modellare il DW non è applicabile il modello ER standard, ma viene usato il DFM (*Dimensional Fact Model*). La rappresentazione con il DFM prevede i seguenti costrutti di base:

- *fatto*: è un concetto di interesse per il processo decisionale, modella un insieme di eventi che accadono nell'azienda (costituisce inoltre una relazione molti-a-molti tra le dimensioni);
- *misura*: è un valore numerico relativo ad un fatto e ne esprime un aspetto quantitativo per l'analisi;
- *dimensione:* è una proprietà con un dominio finito di un fatto; ne descrive una coordinata di analisi;
- *attributi dimensionali:* si intendono le caratteristiche che descrivono una data dimensione;
- *gerarchia:* è un albero orientato i cui nodi sono attributi dimensionali; modellano associazioni tra coppie di attributi.

Il DFM prevede inoltre una serie di attributi avanzati, relativamente poco usati, quali convergenze, additività, dimensioni/attributi opzionali, gerarchie condivise, ecc.

La progettazione concettuale avviene partendo dalla documentazione relativa ai dati riconciliati, quali modelli ER, schemi XML, ecc. La progettazione avviene seguendo i seguenti passi:

1. Definizione dei fatti;
2. Per ogni fatto:
   (a) Costruzione dell'albero degli attributi ed editing;
   (b) Definizioni di dimensioni e misure;
   (c) Creazione dello schema di fatto.

L'albero degli attributi corrispondente a F può essere costruito in modo automatico applicando una procedura che naviga ricorsivamente le dipendenze funzionali espresse, nello schema sorgente, dagli identificatori e dalle associazioni a-uno.

Editing dell'albero. Dal momento che non tutti gli attributi sono di interesse per il data-mart, l'albero è manipolabile eliminando alcuni nodi secondo alcune regole di consistenza. La **potatura** di un ramo avviene eliminando il nodo radice e tutti i relativi figli, l'**innesto** avviene quando un nodo non è necessario ma serve mantenere i suoi nodi figlio.

Definizione delle dimensioni. Vanno scelte nell'albero degli attributi tra i vertici figli della radice. La loro scelta è rilevante in quanto definiscono la granularità degli eventi primari. Nota: il tempo dovrebbe sempre essere una dimensione.

6.1.1 Carico di lavoro e volume dati

Il carico di lavoro di un sistema OLAP è estemporaneo, e va identificato in fase di progettazione, sulla base della reportistica standard e dei colloqui con gli utenti.

Una volta desunto il carico di lavoro del DW attraverso un qualche log di sistema, è possibile attuare una fase di *tuning* del sistema, ossia una specie di ottimizzazione.

Problema della sparsità. Il problema della sparsità dei dati è un difetto del modello multidimensionale che consiste nell'avere coordinate su un evento che potrebbe anche non verificarsi. Tutto ciò comporta uno spreco di risorse, ovviamente. Le alternative per ovviare/attenuare il problema sono 2:

- ROLAP: tiene traccia solo degli eventi realmente accaduti;
- MOLAP: riduzione al minimo dello spazio necessario a tenere traccia degli eventi non accaduti.

6.2 Progettazione logica

I modelli logici generali per la progettazione logica sono 2:

- MOLAP (Multidimensional Online Analytical Processing) utilizzano strutture multidimensionali (array multidimensionali), rappresentano soluzioni che non necessitano di istruzioni complesse in SQL, ma non hanno strutture dati standard. Gestisce la sparsità usando tecniche di compressione dei chunk in modo tale da non sprecare spazio per rappresentare dati sparsi;
- ROLAP (Relational Online Analytical Processing) usano semplicemente il modello relazionale. Il modello ROLAP usa uno schema a stella.

Schema a stella. Uno schema a stella è uno schema caratterizzato da:

- Un insieme di relazioni $DT_1, DT_2, \ldots, DT_n$ ognuna delle quali contiene una chiave primaria e il relativo insieme di attributi;
- Una *fact table* FT, che importa le chiavi di tutte le dimension table.

Lo schema a stella è vantaggioso in quanto è necessario un solo join per recuperare le informazioni in tutte le dimensin table, inoltre non ha problemi di sparsità in quanto vengono memorizzate solo le tuple corrispondenti a punti dello spazio per il quale si sono verificati eventi.

Schema snowflake. Lo snowflake schema è una variante dello schema a stella in cui le DT hanno una chiave e un sottoinsieme di attributi che ne dipendono. Inoltre hanno 0 o più chiavi importate da altre DT necessarie alla ricostruzione del contenuto della DT iniziale. La FT contiene solo le chiavi di alcune DT (*DT primarie*), mentre non contiene quelle delle DT secondarie.

I principali vantaggi dello snowflake schema sono la semplificazione delle interrogazioni e la comodità in presenza di dati aggregati, tuttavia è necessario inserire le chiavi surrogate, e il tempo di interrogazione aumenta data la necessità di dover fare join multipli.

6.2.1 Viste

Le viste sono in sostanza le fact table contenenti dati aggregati di sintesi. Si distinguono in 2 categorie:

- *primarie:* corrispondono al pattern di aggregazione primario (non aggregato);
- *secondarie:* corrispondono a pattern di aggregazione secondari (aggregati).

La presenza di fact table multipli pone il problema di costruire la vista che risolve l'interrogazione da fare con il minor costo possibile. A tale scopo sono usati gli *aggregate navigator*, ossia dei software specifici che si occupano di formulare interrogazioni OLAP sulla miglior vista a disposizione.

Il passaggio dal progetto concettuale a quello logico funziona attraverso 4 fasi:

1. Scelta dello schema da utilizzare;
2. Traduzione degli schemi concettuali;
3. Scelta delle viste;
4. Applicazione di altre forme di ottimizzazione (frammentazione).

La traduzione da schema di fatto a schema a stella avviene semplicemente creando una FT contenente tutte le misure e gli attributi descrittivi collegati direttamente con il fatto e per ogni gerarchia creare una DT che ne contenga tutti gli attributi.

Scelta delle viste. La scelta delle viste è un task complesso, che deve tenere conto di numerosi fattori, quali:

- Minimizzazione delle funzioni di costo;
- Vincoli di sistema;
- Vincoli dell'utente.

In generale conviene materializzare una vista quando risolve un'interrogazione molto frequente, oppure riduce il costo di esecuzione di molte interrograzioni. Viceversa, non conviene quando il suo pattern di aggregazione è molto simile a quello di altre viste già materializzate, oppure non riduce di almeno un ordine di grandezza il costo dell'interrogazione.

6.3 Progettazione dell'alimentazione.

È la fase in cui vengono progettate le procedure necessarie al caricamento dei dati nel data mart.

Estrazione dei dati. In particolare, l'estrazione dei dati può essere effettuata:

- In maniera assistita dall'applicazione;
- Tramite log;
- Tramite trigger;
- Tramite marche temporali.

In ogni caso, i dati estratti consistono di tutti i record che non sono stati estratti nell'ultima operazione di estrazione, e vengono messi nella staging area.

Caricamento dei dati. La modalità di caricamento dei dati nella staging area dipende dal tipo di estrazione e dalla storicizzazione del database. Se l'estrazione è completa, allora la riscrittura del data mart sarà totale; se l'estrazione è incrementale si salva anche il tipo di operazione che ha determinato la variazione del dato. In presenza di storicizzazione si salva anche una coppia di marche temporali indicanti l'intervallo di validità della tupla.

Pulizia dei dati. È l'insieme delle operazioni atto a fixare le incompatibilità tra i vari dati, quali errori di battitura, incompatibiltà di formati, inconsistenza di valori, ecc. Le tecniche utilizzate spaziano dai dizionari (tabelle di look-up) , join approssimati o regole di dominio specifiche.

Alimentazione delle table. L'alimentazione delle dimension table viene svolta semplicemente trasformando gli identificatori in chiavi surrogate.