

La qualità dei dati

E' fuori di dubbio che il successo delle aziende sia sempre più legato alla raccolta e all' utilizzo di grandi quantità di informazioni.

Le decisioni a tutti i livelli sono guidate inevitabilmente dai dati acquisiti da un numero sempre maggiore di fonti e fruiti attraverso svariate tipologie di sistemi (datawarehouse, CRM, ERP) per i quali gli investimenti rappresentano una quota significativa del budget aziendale.

La scarsa qualità dei dati, problema solo in parte percepito come "affrontabile" e "gestibile", determina danni in termini di extra costi, mancati ricavi, demotivazione all'interno dell'azienda.

Quello della qualità dei dati è, in effetti, un aspetto che pochissime aziende percepiscono determinante per i loro livelli di servizio, i loro ricavi e i loro costi. Eppure dovrebbe risultare evidente che c'è un chiaro problema di qualità dei dati alla base della perdita di un cliente che, ad esempio, arriva in un albergo e scopre che la sua prenotazione manca dal database e non c'è ulteriore disponibilità di camere (a chi non è accaduto almeno una volta?).

La nostra esperienza nello svolgere attività di consulenza per le aziende ci ha portato a constatare che nella maggior parte dei casi la insufficiente qualità dei dati viene confusa con un problema di scorretto funzionamento dell'informatica e molte aziende intervengono sui sistemi informativi nella convinzione di risolvere così il problema. Nessuna azienda fornitrice di soluzioni software, in effetti, suggerisce ad un potenziale acquirente di un sistema informativo che la soluzione può essere invece quella di definire ed attuare una strategia di assicurazione e di controllo della qualità dei dati. Ce ne rendiamo conto quando i nostri dati continuano ad essere inesatti anche con il nuovo sistema.

Ma perché la qualità dei dati è così importante? La risposta arriva da Thomas C. Redman, uno dei mostri sacri del settore, che elenca, tra gli altri, due punti essenziali:

La scarsa qualità dei dati è pervasiva

Per capirlo usiamo un esempio. Ci rivolgiamo ad un' agenzia turistica per organizzare un viaggio che prevede l'utilizzo, una volta arrivati sul luogo di destinazione, di un'auto a noleggio per recarsi dall'aeroporto fino al luogo in cui abbiamo deciso di trascorrere le nostre vacanze.

Arrivati all'aeroporto, ci rechiamo presso lo sportello dell'azienda che deve fornirci l'auto e scopriamo, circondati dai nostri bagagli, che la nostra auto non è disponibile e che, come sempre avviene nei periodi di alta stagione, non ve ne sono altre disponibili prima delle prossime 48 ore. Il motivo è che la prenotazione dell'auto non risulta nel database. La persona preposta alla consegna delle auto non sa spiegarsi l'accaduto e telefona all'agenzia che aveva organizzato il viaggio cercando informazioni sulla nostra prenotazione. Dopo vari minuti di telefonata, l'agenzia si rende conto che per un errore, la prenotazione dell'auto non era stata inserita nei dati del viaggio e che, quindi, l'azienda di noleggio auto, non poteva averla ricevuta.

A questo punto si scatena una serie di eventi:

1. Noi telefoniamo irritati all'agenzia
2. La persona dell'agenzia che risponde è mortificata per l'accaduto e promette che ci troverà un'altra auto nel giro di un'ora.
3. Inoltre, l'agenzia ci offre di pagare l'auto per scusarsi dello spiacevole accaduto.
4. Nonostante l'auto a disposizione gratuitamente, continuiamo ad essere seccati per quanto è successo e non esitiamo a raccontare il tutto ad amici e conoscenti.
5. Decidiamo di non rivolgerci mai più a quella agenzia di viaggi

Un banale errore di immissione dati in un database ha quindi determinato per l'agenzia:

1. Lo sconforto dell'impiegato
2. Il costo da sostenere per pagare l'auto a noleggio che ci è stata data gratuitamente
3. La perdita sicura di un cliente (noi)
4. Un numero imprecisato di mancati clienti (i nostri conoscenti e amici ai quali abbiamo raccontato l'accaduto). Quindi, una quantità imprecisata di mancati ricavi.

Questo esempio, analogo a quello esposto da Redman nel suo libro "Data Quality for the Information Age" è, pur nella sua banalità, estremamente significativo sul ruolo che una strategia per l'assicurazione e il controllo di qualità dei dati può avere in azienda. Le ripercussioni in termini di costi e di mancati ricavi generate dalla scarsa attenzione ai dati possono essere veramente devastanti.

La scarsa qualità dei dati è costosa

Il dato fornito dal Datawarehousing Institute parla di una spesa, negli Stati Uniti, di 600 miliardi di dollari all'anno determinata dalla non gestione dei dati. Lettere inutili, telefonate evitabili, staff sovradimensionati. Introdurre in azienda processi di gestione della qualità dei dati viene percepito come un costo che spesso scoraggia i decisori. Il motivo è che manca la percezione dei costi indotti dalla scarsa qualità dei dati. Molti di tali costi, infatti, sono in qualche modo nascosti, nel senso che non sempre vengono associati alla insufficiente qualità dei dati: basti pensare al costo di un depliant pubblicitario che verrà inviato ad un indirizzo inesatto e che andrà, quindi, inevitabilmente perso. Oppure si pensi al tempo che il personale addetto ad un Call Center impiega per discutere con un cliente insoddisfatto. Gli esempi sono molteplici. E quanto costa all'azienda una decisione presa sulla base di informazioni errate?

La qualità dei dati degrada con il tempo

Secondo un'indagine del Datawarehousing Institute (*Data Quality Survey 2001*), le motivazioni che portano degrado alla qualità dei dati sono le seguenti:

- Data entry del personale (76% delle aziende intervistate)
- Data entry dei clienti (25% delle aziende intervistate)
- Modifiche ai sistemi/applicazioni (53% delle aziende intervistate)
- Migrazioni/Conversioni di dati (48% delle aziende intervistate)
- Errata comprensione delle aspettative degli utenti (46% delle aziende intervistate)
- Afflusso di dati esterni (34% delle aziende intervistate)
- Errori di sistema (26% delle aziende intervistate)
- Altri (12% delle aziende intervistate)

Bisogna inoltre considerare che la dinamicità dei dati su cui operiamo ha raggiunto livelli tali che applicare una strategia di controllo di correttezza limitata alla fase di acquisizione dei dati stessi non può più proteggerci dall' avere amare sorprese dopo breve tempo. Oggi quasi tutto è dinamico: indirizzi, numeri di telefono, professioni, indirizzi e-mail; senza parlare dei dati che sono tradizionalmente dinamici: dati di mercato, giacenze di magazzino e così via. In una parola, l'obsolescenza è uno dei nemici più pericolosi; Risulta quindi evidente che la gestione della qualità dei dati deve diventare una attività costante in ogni tipologia di azienda.

Ma cos' è realmente la qualità dei dati?

La definizione più comune deriva dalla constatazione che un qualsiasi database non è altro che una rappresentazione di un qualche aspetto del mondo reale. In altre parole, un database modella una realtà; il nostro problema è fare in modo che la modellazione risulti, nel tempo, il più fedele possibile. Partiamo quindi dalla seguente definizione:

La "Qualità dei Dati" :

- E' il livello di rispondenza delle basi informative alle realtà modellate
- E' il livello di efficienza con il quale i dati servono gli obiettivi di business

Si tratta in effetti di un concetto bidimensionale nel quale vediamo che la prima delle due dimensioni si riferisce alla qualità della relazione tra la base dati e la realtà che essa modella in un dato istante di tempo (la realtà evolve). Esistono alcuni indicatori che ci aiutano nel valutare la qualità con cui una realtà viene modellata da un database:

La completezza delle entità

Indica la differenza che esiste tra la cardinalità del database e quella della realtà modellata

La completezza degli attributi

Ci dice quanto completamente ciascuna entità della realtà è rappresentata nel database

Accuratezza

E' un parametro legato alla correttezza dell'informazione. Il grado di accuratezza può essere definito come la distanza tra il valore riportato nel database e il corrispondente valore reale. L'accuratezza viene misurata a livello di attributi.

Consistenza

La consistenza si riferisce agli attributi. In particolare ne indica la compatibilità tra i valori. Un tipico caso di inconsistenza è quello che può esistere tra CAP e Città o tra Città e prefisso telefonico.

Validità

Indica il livello di aggiornamento degli attributi dinamici.

La seconda delle due dimensioni fa invece riferimento al contesto in cui i dati vengono utilizzati. Anche in questo caso facciamo riferimento a due parametri o indicatori :

Disponibilità

Indica la prontezza con cui il dato è disponibile a coloro che ne hanno bisogno. In effetti è il tempo che intercorre tra l'insorgere dell'esigenza del dato e la sua effettiva disponibilità

Accessibilità

Indica la facilità con cui i fruitori hanno accesso ai dati di cui hanno bisogno.

Gli indicatori che abbiamo elencato, una volta valorizzati, offrono la possibilità di determinare il livello di fiducia attribuibile ai dati che utilizziamo. Già, perché qualunque decisore deve potersi fidare dei dati sui quali prende le sue decisioni o, quantomeno, deve essere in grado di capire quanto può fidarsi dei dati.

Il sistema di gestione della qualità dei dati

In sintesi la qualità dei dati può essere misurata mediante due gruppi di indicatori: il primo che di fatto rappresenta la qualità con cui il database viene alimentato e gestito, il secondo che rappresenta la qualità con la quale i fruitori dei dati vengono agevolati nell'accesso al database.

Per coloro che ritengono la qualità dei dati strategica per il loro business, il problema consiste quindi nel costruire un sistema di assicurazione e di controllo fatto di processi e di risorse (umane e tecnologiche). Tale sistema deve far sì che i parametri o indicatori elencati in precedenza vengano valorizzati, analizzati e gestiti in modo che i corrispondenti valori rientrino costantemente nei limiti pre-fissati.

Ma affinché il sistema di gestione della qualità dei dati sia efficiente ed efficace, l'organizzazione del sistema stesso deve partire dai vertici aziendali. Il tema "qualità dei dati" deve essere una voce permanente nelle riunioni di vertice sia che si tratti di riunioni dei dirigenti sia che si tratti del consiglio di amministrazione. La gestione della qualità dei dati deve entrare a far parte della cultura aziendale. Vediamo come.

Il primo passo, per le aziende che mai hanno affrontato il problema, sarà quello di istituzionalizzare il ruolo di "responsabile della qualità dei dati", ruolo che avrà la responsabilità del corrispondente sistema di gestione e che avrà come primo compito quello di svolgere una attività di diagnosi e, successivamente, di stilare un piano di azione (Il Piano di Qualità dei Dati). In particolare i passi essenziali sono:

- L'identificazione dei dati critici per l'azienda: sono critici quei dati che possono determinare decisioni e/o azioni errate e quindi dannose per l'azienda
- La determinazione dell'attuale livello di fiducia attribuibile ai dati indicati come critici
- La definizione della strategia per aumentare la fiducia nei dati critici : Il piano operativo
- Il piano degli investimenti e la stima del ROI (Return On Investment) associato

E' facile immaginare come l'esecuzione di queste attività non possa non richiedere il coinvolgimento di tutta l'azienda. In particolar modo, per identificare i dati critici sarà necessario analizzare l'intera catena del valore ricostruendo i flussi di dati inter e intra processo, esaminando l'uso che ne viene fatto e le decisioni che ne sono influenzate. E' ovvio quindi che un'analisi di questo tipo non possa essere svolta totalmente dal responsabile della qualità dei dati, che avrà invece bisogno della collaborazione degli attori coinvolti nei processi aziendali. D'altra parte il responsabile avrà il compito di condurre l'analisi, di valutarne i risultati e di confrontarsi con il vertice aziendale. Il risultato della prima attività sarà quindi l'elenco dei dati considerati critici per l'azienda e che devono quindi essere l'oggetto della strategia di assicurazione e controllo di qualità.

Il secondo passo sarà quello di svolgere un'analisi per sapere qual è l'attuale livello di fiducia attribuibile ai dati identificati come critici. Per questo possiamo utilizzare gli indicatori elencati in precedenza.

La valorizzazione degli indicatori ci permetterà quindi di capire quanto siamo lontani dai livelli di qualità necessari e, di conseguenza, ci permetterà di impostare la strategia di qualità ed organizzarla in un piano operativo in cui viene definito nel dettaglio il nuovo sistema di gestione della qualità dei dati.

E' ovvio infine, che il piano operativo dovrà essere accompagnato da un conto economico che mostri costi e benefici.

Lo stato dell'arte

Purtroppo, come abbiamo visto, il tema "Qualità dei Dati" non viene percepito come un aspetto essenziale da curare. In effetti, tutte le aziende soffrono di scarsa qualità dei dati, ma poche sono consapevoli della possibilità che esiste di isolare il problema e di risolverlo senza dovere sostituire il sistema informativo. Inoltre, intervenire per migliorare la qualità dei dati contribuisce alla efficacia di investimenti consistenti nell'acquisizione di sistemi per la fruizione delle informazioni; che si tratti di ERP, CRM o di sistemi per il Datawarehouse. Vorremmo enfatizzare questo aspetto che riteniamo essenziale: l'acquisizione dei succitati sistemi è, in termini di investimento, estremamente impegnativa per qualsiasi azienda; il rischio di investimento è tanto maggiore quanto minore è la consapevolezza del livello qualitativo dei dati che

alimenteranno tali sistemi. Tanto per citare un esempio: di che valore può essere per un'azienda un sistema per il datawarehouse alimentato da dati dei quali non conosciamo il livello di qualità?

Ma torniamo allo stato dell'arte. Ad oggi il mercato delle soluzioni offre alcune tipologie di strumenti classificabili in:

- Data parsing
- Data cleansing
- Record matching

Si tratta di strumenti che permettono di valutare il livello qualitativo dei dati e, ma solo in parte, di migliorarlo. In particolare:

Strumenti di data parsing

Sono strumenti che intervengono sulla struttura sintattica del database. Un esempio chiarirà meglio il concetto:

Se nel nostro database il “**Cliente**” è definito semplicemente come

Nome : char (40)

Indirizzo : char (120)

risulta ovvio che il nostro database conterrà record come il seguente:

Nome: Paolo Rossi

Indirizzo: Via Montegrappa 25 56100 Pisa

Una situazione come questa presenta diversi problemi. In particolare il fatto che il sistema non è totalmente in grado di filtrare alcune tipologie di possibili errori in fase di data entry. Ad esempio se scrivessimo

Indirizzo: Via Montegrappa 2w 5610g Pisa

il sistema non avrebbe modo di segnalarci gli errori evidenti che abbiamo commesso. Uno strumento di data parsing, pur non risolvendo completamente il problema, riduce il range di possibili errori che possiamo commettere in quanto, analizzando i record, trasformerebbe la struttura del “**Cliente**” in:

Nome: char (20)

Cognome: char (20)

Via: Char (50)

No. Civico: int

CAP:int

Città: char (40)

Questa tipologia di strumenti va, tuttavia, utilizzata con estrema cautela in quanto avendo impatto sulla struttura delle entità potrebbe determinare la necessità sulle applicazioni che ne fanno uso.

Strumenti di data cleansing

Si tratta di strumenti che svolgono un'attività di confronto tra il database e la realtà (o quanto meno una porzione di essa) che esso modella.

Se avessimo un record:



Nome: Paolo
Cognome: Rossi
Via: Monterappa
No. Civico: 29
CAP: 56100
Città: Pisa

uno strumento di data cleansing rileverebbe, che Via Monterappa non esiste nello stradario della città di Pisa e proporrebbe di sostituirlo con "Montegrappa". In effetti le proposte del sistema verrebbero scelte dallo stradario secondo il principio della prossimità sintattica. Per cui se nello stradario fosse presente anche Via Montegrappa, questa soluzione verrebbe senz'altro proposta. Questo per dire che lo strumento è sicuramente in grado di rilevare errori certi, ma deve essere in qualche modo assistito nella scelta delle correzioni.

Strumenti per il record matching

Si tratta di strumenti particolarmente utili in quanto permettono di identificare e di rimuovere duplicati eventualmente presenti nel database. Un esempio:

Nome: Paolo
Cognome Rossi
Via: Claudio Monteverdi
No. Civico: 29
Città: Pisa
CAP: 56100

Nome: Paolo
Cognome: Rossi
Via: C. Monteverdi
No. Civico: 29
Città: Pisa
CAP: 56100

Due record come quelli mostrati rappresentano con altissima probabilità la stessa entità (cliente nel nostro caso). Questo vuol dire che se non svolgessimo alcuna attività di rimozione dei duplicati, il Sig. Rossi riceverebbe sempre la stessa comunicazione due volte. Gli strumenti di record matching permettono appunto di rilevare questa tipologia di situazioni e di rimuoverle.

Un elenco di strumenti di gestione della qualità dei dati può essere reperita all'indirizzo :

http://www.ctg.albany.edu/publications/reports/data_quality_tools/data_quality_tools.pdf

Dall'elenco si nota tuttavia come tali strumenti siano estremamente costosi e la loro acquisizione quindi deve essere ben ponderata ; in particolare sconsigliamo vivamente di procedere all'acquisto di tali strumenti al di fuori di una chiara strategia di gestione della qualità dei dati. Mentre riteniamo che possano rappresentare un ottimo investimento se il loro utilizzo è parte integrante di un sistema definito di assicurazione e di controllo della qualità dei dati.

Conclusioni

Fidarsi dei propri dati è un obiettivo irrinunciabile per qualsiasi tipologia di azienda. Occuparsi della qualità dei dati sui quali prendiamo quotidianamente decisioni strategiche è una attività che deve entrare nella lista dei processi cosiddetti “di supporto”.

Per concludere vi proponiamo un breve questionario. Se le risposte che vi darete susciteranno in voi preoccupazione, allora è arrivato il momento di affrontare il problema.

- *Si sono verificati danni o perdite all'interno dell' azienda a causa della scarsa qualità dei dati?*
- *Nei prossimi due anni, quale percentuale di ricavi dipenderà da decisioni e processi automatici basati sui nostri dati elettronici?*
- *Nelle riunioni di management e nei consigli di amministrazione stiamo dedicando sufficiente attenzione ai problemi legati ai dati?*
- *Chi è il responsabile ultimo della qualità dei nostri dati?*
- *Abbiamo una strategia definita per la gestione dei dati?*
- *Ci fidiamo della qualità dei nostri dati?*

Marcello Sabatini

Via dell'Occhio 12

56125 Pisa

marcello.sabatini@tin.it

Tel. 050-2200167

Cell. 335-5353052

Web: www.msconsulting.it , www.datatrust.it