UNIVERSITA' POLITECNICA DELLE MARCHE

FACOLTA' DI INGEGNERIA

Corso di Laurea Triennale in Ingegneria Informatica e dell'Automazione

Data Governance: controllo qualità dei dati nei progetti informatici

Data Governance: Data Quality Control in IT Projects

Relatore:

Prof.ssa **Claudia Diamantini**

Tesi di Laurea di:

Albana Bardho

Matricola: 1031672

A.A. **2021 / 2022**

“a mio figlio”

Indice

Introduzione

La qualità dei dati rappresenta un aspetto molto critico per qualsiasi tipo di progetto. La capacità di produrre una buona analisi e applicare delle scelte nelle varie fasi dello sviluppo ed elaborazione dipende enormemente dalla qualità dei dati utilizzati. Ne consegue che la loro gestione, non è uno step così semplice da attuare richiedendo accorgimenti precisi nel momento della prevenzione, del rilevamento e durante la risoluzione dei problemi in tutte le fasi di un qualsiasi progetto. In merito a ciò si è deciso di investigare di più il ruolo della "data governance" e della qualità dei dati nei progetti informatici.

La tesi sarà strutturata nel seguente modo:

- Il primo capitolo parlerà delle basi di dati soffermandosi in particolare sulle definizioni di sistemi informativi/informatici e dei sistemi di dati necessari per l'approccio del problema.
- Nel secondo capitolo si spiegheranno le caratteristiche generali del data management
- Nel terzo capitolo verrà trattata la data governance e la sua importanza in qualsiasi processo soffermandosi sullo stato dell'arte.
- Nel quarto capitolo, infine verrà trattata l'applicazione della data governance nei progetti informatici con particolare attenzione alle metodologie di miglioramento della qualità dati.

1. I Dati e Basi di Dati

1.1 Dati e Produzione dell'Informazione

Un dato (dal latino "datum" che significa letteralmente fatto) è una descrizione elementare di un'entità, di un fenomeno, di un avvenimento o di altro.
I dati nascono dell'osservazione di aspetti e permettono di effettuare così operazioni numeriche, risolvere un problema, caratterizzare un fenomeno o di esprimere un'opinione.
L'elaborazione dei dati può portare alla conoscenza di un'informazione. A questo punto è necessario dare una definizione precisa anche di questi due concetti appena citati:

- I dati sono rappresentazioni originarie, cioè non interpretate, di un fenomeno, evento, o fatto, effettuate attraverso simboli o combinazioni di simboli, o di qualsiasi altra forma espressiva legate a un qualsiasi supporto.
- L'Informazione è una visione della realtà derivante dall'elaborazione e interpretazione dei dati. insieme di dati comprensibili per il destinatario (Di Nunzio 2014)

In conclusione, la relazione fra dato, elaborazione ed informazione la si può riassumere con la seguente rappresentazione (Fig.1):

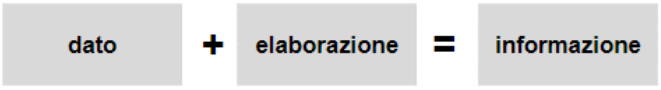

Figura 1 Produzione dell'informazione

In generale il processo di produzione delle informazioni si può definire in tre fasi:

- acquisizione dei dati (elementari);
- elaborazione dei dati;
- emissione dell'informazione.

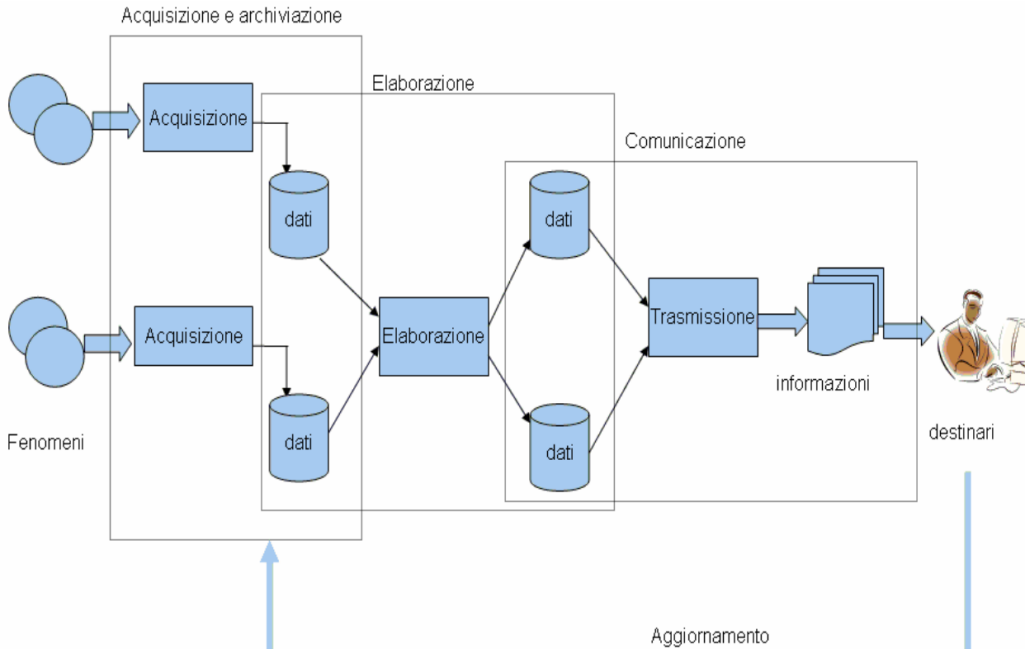

Figura 2 Produzione dell'informazione

Il processo può essere sintetizzato e descritto in figura 2. In definitiva, si può concludere che le informazioni di un'organizzazione sono disponibili sotto forma di un insieme di dati memorizzati su un apposito apparato.

Inoltre, i dati devono essere opportunatamente 'interpretati' per dare luogo alle informazioni vere e proprie.

Sia i dati "grezzi" ("raw data") sia le regole per la loro interpretazione e per poterli immagazzinare sono memorizzati sotto forma di dati. Ne consegue che essi sono immediatamente presenti alla conoscenza prima di ogni elaborazione. I dati opportunatamente interpretati forniscono informazione.

1.2 Sistema Informativo e Sistema Informatico

In un qualsiasi tipo di organizzazione, l'informazione è una risorsa significativa al pari di altri tipi di risorse: come quelle umane o materiali.
Con il termine "Organizzazione" si intende un insieme di uomini, strumenti, attività coordinato per il raggiungimento di obbiettivi comuni. A questo punto risulta necessario definire cosa sia un sistema informativo ed informatico.
Il sistema informativo di un'organizzazione può essere definito come una combinazione di risorse e di procedure organizzate per:

- la raccolta
- l'archiviazione
- l'elaborazione
- lo scambio delle informazioni necessarie alle attività operative alle attività di programmazione e controllo (informazioni di gestione), e alle attività di pianificazione strategica (informazioni di governo).

Mentre, in maniera opposta un sistema informatico è la tecnologia che fornisce supporto al sistema informativo, cioè: Macchine hardware, Programmi software, Banche dati e sistemi di gestione e Reti di comunicazione (Chianese, et al. 2015)
Concludendo possiamo dire che i sistemi informatici si basano sull' informatica per il trattamento dei dati e la produzione delle informazioni. Le procedure sono automatizzate e costituite da programmi funzionanti su calcolatore.

1.3. Sistemi per Basi di Dati

Una base di dati è un insieme di informazioni associato a collezioni di dati, correlati fra di loro (Chianese, et al. 2007).
Si tratta quindi di un unico e grande deposito di dati che può essere condiviso all'interno dell'azienda da tutte le applicazioni che ruotano all' interno.
Esso risulta "persistente" cioè con vita molto più lunga delle procedure di gestione interne, il che consente di lavorare sempre su uno stato consistente dei dati.
I Dati permanenti raccolti e gestiti da un elaboratore elettronico sono suddivisi in due categorie distinte:

- I **metadati**, ovvero lo schema della base di dati. Si tratta di definizioni che descrivono la struttura dei dati, le restrizioni sui valori ammissibili dei dati (vincoli d'integrità). (Chianese et al. 2007)

  I metadati vengono definiti prima di creare i dati ed indipendentemente dalle applicazioni che usano la base di dati.

- I **dati** hanno le seguenti caratteristiche:

- Essi risultano organizzati in insiemi omogenei, fra i quali sono definite delle relazioni. La struttura dei dati e le relazioni sono analizzate e descritte nello schema con opportuni meccanismi dipendenti dal modello dei dati (data model) utilizzato, che prevede anche operatori per estrarre elementi da un insieme e per conoscere quelli che, in altri insiemi, sono in relazione con loro (Albano et al 2021),
- Essi sono permanenti, cioè, una volta generati, continuano ad esistere finché non ´sono esplicitamente rimossi; la loro vita, quindi, non dipende dalla durata delle applicazioni che ne fanno uso;
- Risultano accessibili mediante transazioni ("transactions"), unità di lavoro che non possono avere effetti parziali
- sono protetti sia dall'accesso di utenti non autorizzati, sia da corruzione dovuta a malfunzionamenti hardware e software
- sono utilizzabili contemporaneamente da utenti diversi (Albano et al 2021)

2. Data Management

2.1. Sistemi per la gestione di Base di Dati: DBMS – Data Base Management System

Un DBMS (Data Base Management System) è un sistema software, centralizzato o distribuito, il quale consente di poter applicare le seguenti funzioni:

- Definire specifici schemi di basi di dati,
- Scegliere le strutture dati per la memorizzazione e l'accesso ai dati
- Memorizzare, recuperare e modificare i dati, interattivamente o da programmi, ad utenti autorizzati e rispettando i vincoli definiti nello schema.

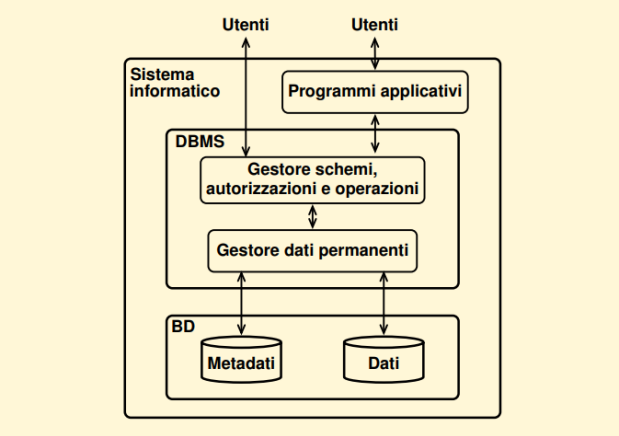

Figura 3. Il sistema informatico e il sistema per la gestione di basi di dati (Albano, Ghelli e Orsini 2021)

Le proprietà rilevanti di una base di dati sono:

- Integrità.

  I DBMS prevedono dei meccanismi per controllare che i dati inseriti, o modificati, siano conformi alle definizioni date nello schema, in modo da garantire che la base si trovi sempre in uno stato consistente.

  Il linguaggio per la definizione dello schema logico consente di definire non solo la struttura dei dati, ma anche le condizioni a cui essi devono sottostare per essere significativi (vincoli d'integrità), quando queste condizioni vanno verificate e cosa fare in caso di violazioni.

- Affidabilità.

  I DBMS devono disporre di meccanismi e sistemi per proteggere i dati da malfunzionamenti hardware o software e da interferenze indesiderate dovute ad accessi concorrenti ai dati da parte di più utenti.
- Sicurezza.

  I DBMS prevedono meccanismi sia per controllare che ai dati accedano solo persone autorizzate, sia per restringere i dati accessibili e le operazioni che si possono fare su di essi. Il problema presenta diversi aspetti, alcuni dei quali simili a quelli affrontati nell'ambito dei sistemi operativi, altri tipici di questa classe di applicazioni

2.2 Data Quality: Filter e GIGO - "Garbage in – Garbage out"

Le conseguenze della scarsa qualità dei dati sono spesso vissute nella vita di tutti i giorni, ma senza effettuare i necessari collegamenti con le sue cause.
La qualità dei dati è oltremodo riconosciuta come una questione di performance abbastanza rilevante nei processi operativi e nelle attività decisionali (Batini et al 2009).
A tal proposito, il ben noto principio di "Garbage In Garbage Out (GIGO)" indica che indipendentemente da quanto sia avanzato lo sviluppo delle tecniche e delle metodologie rispetto a nuove soluzioni software, la qualità dei dati è ancora un fattore importante nel funzionamento di successo dei sistemi IT (Abdul Aziz, Azwa, Md Yazid Mohd Saman, e Mohd Po 210).
In effetti, una visione tradizionale della modellazione al computer ha perpetuato questo tipo di concetto, che, non importa quanto sia sofisticata l'analisi, i risultati finali potrebbero non essere attendibili. Infatti, se il dato di input non è accurato, l'output non lo sarà altrettanto.

Le dimensioni individuate per identificare la qualità dei dati sono (Batini e Scannapieco 2016):

- “Precisione”: l'imprecisione implica che il sistema informativo rappresenti uno stato del mondo reale diverso da quello che avrebbe dovuto essere “rappresentato”. Imprecisione si riferisce a una mappatura confusa in uno stato errato dell'IS (Information System).
- “Affidabilità” indica “se si può contare sui dati per trasmettere la giusta formazione; Essa può essere vista come la correttezza dei dati”.
- La “tempestività” si riferisce al “ritardo tra un cambiamento dello stato del mondo reale e la conseguente modifica dello stato del sistema informativo”.
- La “completezza” è “la capacità di un sistema informativo di rappresentare ogni stato significativo del sistema rappresentato nel mondo reale”.
- La “coerenza” dei valori dei dati si verifica se esiste più di uno stato del sistema informativo che corrisponde a uno stato del sistema reale; quindi “incoerenza significherebbe che la mappatura della rappresentazione è uno a molti.

In generale si dovrebbero tenere in considerazione queste tre fasi per la qualità dei dati (Batini et al 2009):

1. “State Reconstruction”, che ha lo scopo di raccogliere informazioni contestuali su processi e servizi organizzativi, raccolte di dati e relative procedure di gestione, problemi di qualità e costi corrispondenti; questa fase può essere saltata se delle informazioni sono disponibili da analisi precedenti.
2. “Valutazione/Misurazione”, che misura la qualità delle raccolte di dati lungo le dimensioni di qualità pertinenti.

   Il termine misurazione viene utilizzato per affrontare il problema nel misurare il valore di un insieme di dimensioni della qualità dei dati. Il termine valutazione è utilizzato quando tali misurazioni vengono confrontate con valori di riferimento, al fine di abilitare una diagnosi di qualità.
3. Il “miglioramento” riguarda la selezione dei passaggi, delle strategie e delle tecniche per raggiungere nuovi obiettivi di qualità dei dati

2.3 Decision making process

Il processo decisionale e la sua efficacia sono intuitivamente correlati all'informazione e alla loro qualità.

Per cui la decisione, tra l'altro, è influenzata dall'accuratezza, completezza delle informazioni disponibili.

Con il miglioramento delle tecniche e delle tecnologie di gestione dei dati, i dati continuano a diventare sempre più importanti per le aziende. Un numero crescente di aziende utilizza i dati per prendere decisioni.

Affinché i dati siano utili, tuttavia, devono essere di alta qualità. Anche le nuove tecnologie stanno aumentando l'importanza dei dati e della loro qualità.

Tecnologie come l'intelligenza artificiale e l'automazione hanno un potenziale enorme, ma il successo con queste tecnologie dipende fortemente dalla qualità dei dati.

L'apprendimento automatico, ad esempio, richiede grandi volumi di dati accurati. Più dati validi ha un algoritmo di apprendimento automatico, più velocemente può produrre risultati e migliori saranno i risultati.

Keller e Staelin (1987) hanno indagato gli effetti sia della qualità sia della quantità di informazioni in relazione al “decision making”. Le loro conclusioni possono essere così riassunte:

- L'efficacia delle decisioni è influenzata negativamente dall'aumento della quantità di informazioni messe a disposizione e favorita da incrementi del livello qualitativo, almeno fino a un certo punto.
- L'efficacia delle decisioni prima aumenta e poi diminuisce all'aumentare della quantità di informazioni disponibili, mantenendo fisso il livello medio di qualità dell'ambiente informativo.
- Livelli più elevati di IQ (Information Quality) sono associati a quantità relative più elevate di informazioni sugli “attributi” utilizzati, mantenendo così fissa la quantità di informazioni disponibili.

  Collegando l'uso delle informazioni sugli attributi all'accuratezza della scelta, sono emerse maggiori probabilità di raggiungere la scelta corretta quando venivano utilizzate la maggior parte, ma non tutte le informazioni disponibili.

3. Data Governance

Per governance dei dati si intende l'approccio strategico con cui un'organizzazione stabilisce come gestire i propri dati.

La data governance, quindi, può essere definita come la strategia che, in base agli obiettivi dell'azienda, stabilisce il modello di data management da applicare ai dati.

Secondo Otto (2011) "La governance dei dati è definita come un quadro a livello aziendale per l'assegnazione dei diritti relativi alle decisioni e doveri per poter trattare adeguatamente i dati come patrimonio aziendale".

In maniera opposta, il data management è l'insieme delle tattiche, azioni, strumenti e procedure impiegati per porre in atto le politiche di data governance e far fluire i dati all'interno dell'azienda.

3.1 Introduzione alla Data Governance

In generale con la digitalizzazione, i processi delle imprese e i loro rapporti con i clienti si sono evoluti e producono sempre più dati.

Quando i volumi e la quantità dei dati aumentano fino a diventare i cosiddetti "Big Data", i normali processi e strumenti di gestione non sono più sufficienti e diventa necessaria una vera e propria strategia o metodologia per la gestione della qualità dei dati.

Da un punto di vista operativo, possiamo definire la data governance come la base strategica sulla quale costruire l'intero sistema di gestione dei dati.

Nella revisione della letteratura relativo ai quadri di governance dei dati, il quadro proposto da (Khatri e Brown 2010) è stato selezionato per presentare i domini decisionali che dovrebbero essere considerati per la data governance. Il quadro contiene cinque domini decisionali correlati (Figura 4):

1. Data Principle
2. Qualità dei dati
3. Metadati
4. Accesso ai dati
5. Ciclo di vita dei dati

Ciascuno dei cinque domini decisionali affronta un insieme di problemi core che sono spiegati di seguito

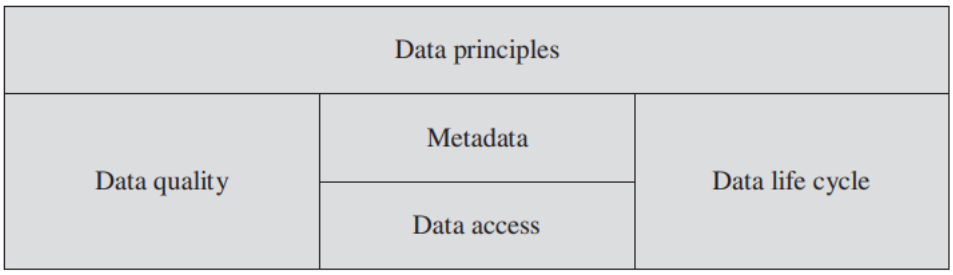

Figura 4 Data Governance domini (Data governance activities: an analysis of the literature)

I principi dei dati sono mostrati nella parte superiore del quadro di figura 4 ed hanno lo scopo di stabilire la direzione per tutti gli altri domini decisionali. Quindi, i principi fissano i requisiti limite per gli usi delle risorse di dati, che servono ad affrontare gli standard aziendali per la qualità dei dati.

La qualità dei dati affina quindi la base per come i dati sono interpretati (metadati) e accessibili (accesso ai dati) dagli utenti.

Infine, la decisione sul ciclo di vita dei dati definisce la produzione, la conservazione e il ritiro delle risorse di dati che vengono riprodotte un ruolo fondamentale nell'operazione dei principi dei dati nell'infrastruttura IT.

I fattori critici di successo per la governance dei dati possono essere determinati prendendo in considerazione alcuni punti principali individuati da Marinos (2004).

Essi sono:

- Responsabilità e responsabilità strategica.

È necessario che la leadership esecutiva guidi il processo di governance dei dati. Per implementare con successo la governance dei dati, i ruoli e le responsabilità delle varie persone nelle organizzazioni coinvolte nel processo di governance dei dati devono essere chiaramente definiti.

- Standard.

La definizione degli standard dei dati è importante in quanto i dati aziendali devono essere definiti e assicurati che siano "adatti allo scopo".

- Abbracciare la complessità.

Gli stakeholder dei dati sono i produttori e i consumatori dei dati. La gestione degli stakeholder dei dati è complessa in quanto i dati possono essere raccolti, arricchiti, distribuiti, consumati e mantenuti da diversi stakeholder dei dati.

- Questione interdivisionale.

La struttura di governance dei dati deve essere progettata in modo tale da includere la partecipazione di tutti i livelli dell'organizzazione per conciliare le priorità, accelerare la risoluzione dei conflitti e incoraggiare il supporto della qualità dei dati.

- Metriche.

La definizione di metriche di qualità dei dati specifiche per i risultati è importante per misurare il successo della governance dei dati.

- Collaborazione.

Quando un'organizzazione condivide i dati con altre organizzazioni (partner), è necessario che il suo partner sia ritenuto responsabile della qualità dei dati in modo che gli sforzi di gestione dei dati di entrambe le organizzazioni non siano compromessi.

- Scelta dei punti strategici di controllo.

È necessario mettere in atto controlli per determinare dove e quando la qualità dei dati deve essere valutata e affrontata.

- Monitoraggio della conformità.

I dati devono essere controllati per essere gestiti. Le attività di controllo sono procedure e tecniche che devono essere in linea con gli obiettivi del progetto e per la mitigazione del rischio. Queste attività di controllo producono alla fine dei report da parte dell'azienda al fine di misurare il raggiungimento di determinati obiettivi prefissati.
I sistemi di controllo interno a tal proposito sono un valido supporto per assicurare i principi di gestione e amministrazione, l'adeguatezza degli assetti e le procedure organizzative aziendali.

3.2 Stato Dell'arte

Tra i ricercatori si tende ad approvare l'idea che la governance dei dati debba trovare risposte a tre tipi di domande (Otto 2011):

1. Quali decisioni devono essere prese a livello aziendale (in termini di dati)?
2. Quali ruoli sono coinvolti nel processo decisionale?
3. Come sono i ruoli coinvolti nel processo decisionale?

Per quanto riguarda la prima domanda, una serie di conclusioni può essere definita. Le decisioni relative alla governance dei dati fanno riferimento ad alcuni principi fondamentali per la gestione dei dati (l'uso di standard di dati ad esempio), ai requisiti per la qualità dei dati.
La misurazione della qualità dei dati, la gestione dei metadati, i requisiti di accesso ai dati così come la gestione del ciclo di vita sono parte fondamentale del "decision making" a livello aziendale.
Per quanto riguarda la seconda domanda, riferendosi ai ruoli coinvolti nella Data Governance, essi sono i "*data steward*", i "*proprietari di dati*" e "*comitati di dati*".
I "*data steward*" supportano i dipartimenti aziendali nell' utilizzo dei dati desiderati. I "data steward" sono anche responsabile della presa e della valutazione dei requisiti aziendali e dei problemi con dati.
Mentre i data steward rappresentano la funzione di gestione dei dati di un'impresa, i "proprietari data" appartengono a determinati dipartimenti o divisioni aziendali. Essi specificano i requisiti aziendali sui dati e sulla qualità dei dati.
Il ruolo del proprietario dei dati per quanto riguarda i dati anagrafici dei fornitori, ad esempio, è spesso assegnato al responsabile degli acquisti centrali. Un comitato dati è la centrale decisionale nella governance dei dati.
Essa specifica i principi da utilizzare per i dati in una impresa, e corrisponde ai diversi interessi ed esigenze delle divisioni aziendali (rappresentate dai proprietari dei dati).
Infine, ciò che concerne la terza questione, riferendosi al collegamento di ruoli e aree decisionali, la Data Governance riguarda l'assegnazione di autorità e, conseguentemente, responsabilità.

Ad esempio, l'autorità decisionale in merito all'architettura dei dati potrebbe essere assegnata ai dati comitato, mentre il potere esecutivo è assegnato all'amministratore dei dati.

Per fare ciò, i diagrammi funzionali sono spesso utilizzati per la modellazione. Un diagramma funzionale è una tecnica utilizzata nella progettazione organizzativa mediante il quale i compiti sono collegati con i ruoli interni alla gestione mediante i cosiddetti "authority codes" o codici di autorità (Otto 2011).

4. Data Governance nei progetti informatici

4.1 I progetti informatici

In generale, un progetto secondo la (ISO 21500.) è un insieme di processi unico, composto da attività coordinate e controllate, con date di inizio e di fine, che vengono realizzate per raggiungimento degli obiettivi del progetto.

Gli elementi che attribuiscono il miglior risultato per raggiungere l'obiettivo sono per esempio:

- Accurate stime iniziali;
- Utilizzo di tecniche appropriate;
- Impegno del gruppo;
- Disponibilità delle risorse necessarie
- Controllo puntuale ma non burocratico
- La suddivisione in fasi della gestione del progetto.

Un progetto informatico è un processo che porta alla realizzazione di un prodotto informatico secondo delle specifiche richieste da parte del cliente.

Le motivazioni per la produzione di un software sono numerose e differenti, tuttavia le motivazioni più comuni sono:

1. Soddisfare esigenze specifiche di uno specifico cliente (Software personalizzato e venduto ad un solo cliente);
2. Soddisfare l'esigenza percepita di alcuni gruppi potenziali di utenti (Software standardizzato e commercializzato a più clienti);

3. Soddisfare un'esigenza personale, ovvero un programmatore scrive un piccolo software per svolgere una determinata azione e/o per risolvere un determinato problema;

Le prime due motivazioni tengono conto di sviluppi software di medio/grandi dimensioni che richiedono uno o più team di progetto.

Durante la gestione del progetto informatico si fa uso di modelli che servono per descrivere il problema.
I cinque modelli più conosciuti per la produzione dei progetti informatici sono**:** (Casadei e Teolis 2013)

1. Modello a cascata (waterfall model)

La caratteristica di questo modello (simile ad una cascata) non è tanto costituita dal numero o dalla descrizione delle fasi che ne fanno parte, ma dal fatto che si passa alla fase successiva di progettazione solo quando la precedente è completamente terminata.
Ogni fase del processo deve terminare con un documento formale approvato dall'organo responsabile dell'intero progetto.
In questa maniera facendo eventuali ripensamenti (che implicano rimettere in discussione i risultati di fasi precedenti) hanno un impatto molto rilevante sui tempi e sui costi e quindi sono una eventualità da prevenire con molta cura.
Nella pratica, può accadere che, a seguito di situazioni particolari, si sia tentati di sottovalutare l'importanza di qualcuna di queste fasi e quindi di ometterla; anche questo deve essere assolutamente evitato perché verrebbero meno la produzione e l'approvazione del documento necessario per avviare la fase successiva.

2. Modello a V (V model)

Il modello a V è caratterizzato (tipicamente) da 9 fasi di sviluppo, disposte come sono rappresentate in Figura 5.
Il modello a V è caratterizzato da interazioni anche non locali tra le diverse fasi. In questo modello è possibile osservare possibili dei ricicli non solo alla fase precedente. Inoltre, esiste una fase finale di revisione critica di tutto il progetto

Il nome del modello deriva dal fatto che le nove fasi sono “in qualche modo” simmetriche in maniera tale che da una fase si può riciclare non solo a quella precedente, ma anche a quella alla stessa altezza se le si dispone in uno schema a “V”.

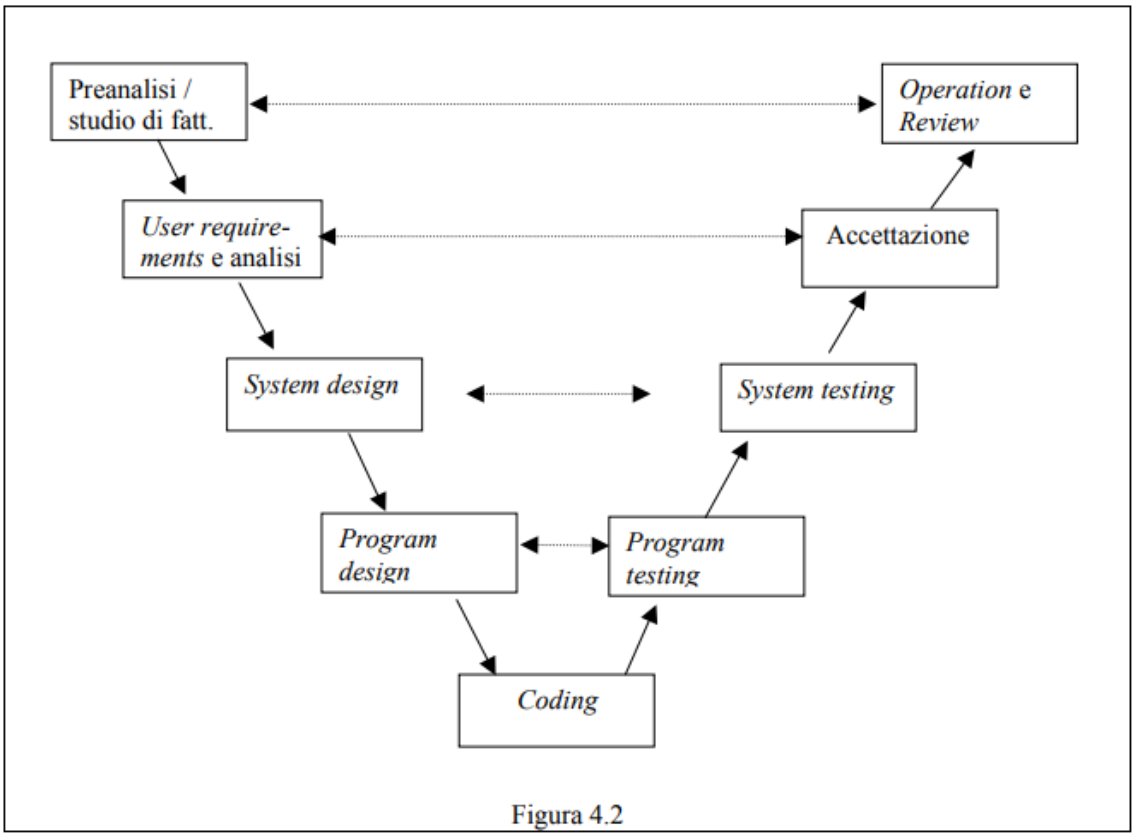

Figura 5 Modello a V (Casadei e Teolis 2013)

3. Modello incrementale (incremental delivery)

Il modello a cascata si basa sui seguenti due presupposti fondamentali:

1. Il committente accetta che tutto l’output del progetto venga messo in esercizio in blocco nell’ultima fase temporale del progetto;
2. i requisiti degli utenti e la tecnologia sono ben noti e accettati sia dal committente sia dallo sviluppatore.

Quando (solo) il secondo presupposto non è valido allora si adotta il modello a V. Quando, invece, non è valido (solo) il primo dei due presupposti, si adotta il modello “a consegna incrementale” che permette al committente di avere, entro tempi brevi, una

prima consegna di risultati, anche se parziali, e consegne successive via via durante lo svolgersi del progetto. I vantaggi di questo modello, a causa delle consegne distribuite nel tempo, derivano dal fatto che il feedback delle prime consegne influenza le successive e quindi gli errori strategici tendono ad evidenziarsi durante lo svolgimento del progetto (piuttosto che solo alla fine).
Gli eventuali errori di analisi tendono ad essere "piccoli" o poco rilevanti perché la consegna è temporalmente vicina alla fase di indagine dei problemi.

4. Modello a prototipi (prototyping model)

I modelli precedenti si caratterizzano per la presenza di entrambi o di uno solo dei seguenti due presupposti fondamentali:

1. Il committente accetta che tutto l'output del progetto venga messo in esercizio in blocco nell'ultima fase temporale del progetto
2. I requisiti degli utenti e la tecnologia sono ben noti e accettati sia dal committente sia dallo sviluppatore.

Ci sono dei casi, caratterizzati dalla incertezza tecnologica e dalla incertezza dei requisiti, in cui occorre necessariamente procedere per tentativi; vale a dire occorre procedere con piccole realizzazioni, coinvolgendo eventualmente un campione di utenti ridotto, per poi discutere su quanto realizzato per mettere a punto le funzionalità necessarie. Si possono adottare due approcci:

1. "Usa e getta": i prototipi che vengono costruiti e usati per ragionare con l'utente, anche in passi successivi, vengono eliminati e sostituiti dal sistema definitivo;
2. "Evolutivo": si parte con un prototipo che viene via via modificato e sviluppato, per approssimazioni successive, fino allo stato di sistema definitivo (cioè accettato dal committente).

5. Modello a spirale (Spiral model).

I modelli visti in precedenza sono prettamente tipici dell'ambiente standard di riferimento dei progetti trattati in questo contesto, cioè dei sistemi informativi delle aziende medio-grandi. In questi ambienti, tipicamente è possibile realizzare progetti "sequenziali", cioè progetti i cui prodotti (studio di fattibilità, requisiti degli utenti, analisi funzionale, analisi di dettaglio, ecc.) sono costruiti, almeno in linea di principio, sequenzialmente.

Ambienti in cui sono di frequente sviluppati progetti software che non sempre consentono uno sviluppo sequenziale sono:

1. applicazioni multimediali (come giochi, enciclopedie, ambienti virtuali);
2. sistemi per il controllo di processi, per esempio in stabilimenti di produzione, raffinerie, impianti per la produzione di energia;
3. sistemi embedded, cioè insiemi di processori e relativo software integrati per il controllo e la gestione di apparecchiature di varia complessità come per esempio: insiemi di elettrodomestici collegati (e "intelligenti"), grosse macchine utensili e soprattutto sistemi militari,
4. sistemi per la simulazione di fenomeni complessi (previsioni meteorologiche, idrodinamica di bacini complessi, sistemi cosmologici, ecc.), • sistemi operativi, software di base, pacchetti applicativi di larga diffusione.

4.2 IT Governance: Planning and monitoring

L'IT Governance è un processo utilizzato per monitorare e controllare le decisioni chiave IT, nel tentativo di dare valore ai componenti di un'organizzazione. L'IT Governance è un processo. Non è un evento puntuale.

L'obiettivo dell'IT Governance è garantire la consegna di risultati aziendali e non solo "prestazioni dei sistemi IT" né tantomeno considerare la "gestione del rischio IT".

Al contrario, l'IT Governance riguarda le decisioni IT che hanno un impatto sul valore aziendale.

Il processo, quindi dell'IT Governance è per certi aspetti monitorare e controllare le decisioni IT che potrebbero avere un impatto - positivo o negativo.

Misurare e controllare la qualità dei dati in una singola organizzazione è un compito complesso. In relazione a ciò, vi è la questione della qualità dei dati nei progetti informatici che è diventata sempre più complessa e controversa in conseguenza dell'evoluzione informatica. Nei sistemi informativi in rete, i processi sono sempre più coinvolti in complessi scambi di informazioni e spesso operano su input ottenuti da fonti esterne, le quali sono spesso sconosciute a priori.

Ne consegue che la qualità complessiva dei dati che fluiscono attraverso i sistemi informativi può degradarsi rapidamente nel tempo se la qualità dei processi e degli input informativi non è controllata.

4.3 Data Quality in IT Projects

I sistemi informativi offrono nuove opportunità per la gestione della qualità dei dati, come la capacità di selezionare e confrontare i dati da diverse fonti per rilevare e correggere gli errori e, quindi, migliorare la qualità complessiva dei dati.

Secondo Bair (2006) la qualità dei dati può essere definita in base a diversi fattori come il loro dominio dei dati, la correttezza, la completezza, unicità e integrità referenziale.

Al fine di determinare che i dati sono "adatti allo scopo", vengono definiti sei dimensioni/caratteristiche quali: la qualità di accuratezza, tempestività, pertinenza, completezza, comprensione e fiducia.

La qualità dei dati è importante per le aziende al fine di sfruttare iniziative IT come data “mining” e “warehousing”.

Il successo di tali investimenti IT dipende molto dalla qualità dei dati di origine. Il concetto "Garbage In, Garbage Out" come già accennato, è molto applicabile in questa situazione.

L'efficacia di qualsiasi iniziativa IT dipende dalla qualità dei dati.

I rapporti generati e le decisioni del prodotto possono essere buoni solo quanto la qualità dei dati.

I problemi che circondano la qualità dei dati o la mancanza di qualità sono aggravati dal fatto che i dati sono diffusi tra sistemi disparati all'interno di un'organizzazione, i dati sono raccolti, mantenuti e utilizzati da vari livelli di un'organizzazione e molti sistemi di sviluppo le metodologie non incorporano la garanzia della qualità dei dati.

I problemi di qualità dei dati sopra menzionati possono essere risolti disponendo di un'efficace gestione dei dati che potrebbe garantire una buona qualità attraverso l'uso di un programma di governance dei dati.

La letteratura fornisce un'ampia gamma di tecniche per valutare e migliorare la qualità dei dati, come il collegamento dei records, le regole aziendali etc..

Nel tempo, queste tecniche si sono evolute per far fronte alla crescente complessità della qualità dei dati nei sistemi informativi in rete.
A causa della diversità e complessità di queste tecniche, la ricerca si è recentemente concentrata sulla definizione di metodologie che aiutano a selezionare, personalizzare e applicare tecniche di valutazione e miglioramento della qualità dei dati. In generale, un'ulteriore suddivisione dei dati è la seguente:

- I Dati strutturati, sono aggregazioni o generalizzazioni di elementi descritti da attributi elementari definiti all'interno di un dominio.
  I domini rappresentano l'intervallo di valori che possono essere assegnati agli attributi e di solito corrispondono a tipi di dati elementari dei linguaggi di programmazione, come valori numerici o stringhe di testo.
  Le tabelle relazionali ei dati statistici rappresentano la tipologia più comune di dati strutturati.
- I dati non strutturati, sono una sequenza generica di simboli, tipicamente codificati in linguaggio naturale. Esempi tipici di dati non strutturati sono un questionario contenente testo libero che risponde a domande aperte o il corpo di un'e-mail.

- I dati semistrutturati sono dati che hanno una struttura che ha un certo grado di flessibilità. Essi vengono anche definiti senza schema o autodescrittivi

La grande maggioranza dei contributi di ricerca nella letteratura sulla qualità dei dati si concentra su dati strutturati e semistrutturati (Batini e et al, 2009).
Nel caso più generale, la sequenza delle attività da tenere in conto per la qualità dei dati è composta solitamente da tre fasi:

1. "Ricostruzione dello Stato", che ha lo scopo di raccogliere informazioni contestuali su processi e servizi organizzativi, raccolte di dati e relative

procedure di gestione, problemi di qualità e costi corrispondenti. Questa fase può essere saltata se le informazioni sono disponibili da analisi precedenti.

2. Valutazione/misurazione, che misura la qualità delle raccolte di dati lungo le dimensioni di qualità pertinenti; il termine misurazione viene utilizzato per affrontare il problema di misurare il valore di un insieme di dimensioni della qualità dei dati.

   Il termine valutazione viene utilizzato quando tali misurazioni vengono confrontate con valori di riferimento, al fine di consentire una diagnosi di qualità.

3. Il miglioramento riguarda la selezione delle fasi, delle strategie e delle tecniche per raggiungere nuovi obiettivi di qualità dei dati.

Nelle loro fasi di miglioramento, le metodologie adottano due tipi generali di strategie, ovvero guidate dai dati e guidate dai processi.

Le strategie basate sui dati migliorano la qualità dei dati modificando direttamente il valore dei dati. Ad esempio, i valori dei dati obsoleti e non più attuali vengono aggiornati con un database più aggiornato.

Le strategie basate sui processi migliorano la qualità riprogettando i processi che creano o modificano i dati.

Le strategie, basate sia sui dati che sui processi, applicano una varietà di tecniche: algoritmi, e attività basate sulla conoscenza, il cui obiettivo è migliorare la qualità dei dati.

Un elenco delle tecniche di miglioramento applicate dalle strategie basate sui dati è:

- Acquisizione di nuovi dati, che li migliora acquisendo dati di qualità superiore per sostituire i valori che sollevano problemi di qualità
- Standardizzazione (o normalizzazione), che sostituisce o integra valori di dati non standard con valori corrispondenti conformi allo standard.
- Collegamento di record, che identifica le rappresentazioni dei dati in due (o più) tabelle che potrebbero riferirsi allo stesso oggetto del mondo reale;

(Olson 2003) ha associato alla scarsa qualità dei dati, l'aumento dei costi e la complessità dello sviluppo della gestione delle relazioni, per esempio, con i clienti.

I costi sono una prospettiva rilevante considerata nelle metodologie, a causa degli effetti di dati di bassa qualità sulle attività che consumano risorse.

Quindi i costi legati ai dati rappresentano un aspetto molto importante sia da un punto di vista dei possibili svantaggi che possono causare, sia dal punto di vista dei miglioramenti necessari che si possono effettuare.

Il costo della qualità dei dati è la somma del costo delle attività di valutazione e miglioramento della qualità dei dati (indicato anche come costo del programma per la qualità dei dati) e il costo associato alla scarsa qualità dei dati.

Di conseguenza, il costo della scarsa qualità può essere ridotto implementando un programma di qualità dei dati più efficace, che è in genere più costoso.

Pertanto, aumentando il costo del programma di qualità dei dati, si riduce il costo della scarsa qualità dei dati. Questa riduzione può essere vista come il vantaggio di un programma di qualità dei dati.

Il costo di un programma per la qualità dei dati può essere considerato un costo preventivo sostenuto dalle organizzazioni per ridurre gli errori nei dati. Questa categoria di costo include il costo di tutte le fasi e passaggi che compongono un processo di valutazione e miglioramento della qualità dei dati.

Di seguito, verranno introdotte alcune metodologie che si possono applicare per il miglioramento della qualità dei dati,

- **La Metodologia DWQ**

La metodologia DWQ è stata sviluppata nell'ambito del progetto European Data Warehouse Quality.

Questa metodologia studia la relazione tra obiettivi di qualità e opzioni di progettazione nel data warehousing. (Batini et al, 2009).

La metodologia considera la soggettività del concetto di qualità e fornisce una classificazione degli obiettivi di qualità, in base al gruppo di stakeholder che persegue tali obiettivi.

La metodologia DWQ afferma che i metadati del data warehouse dovrebbero tenere conto di tre prospettive:

- una prospettiva aziendale concettuale incentrata sul modello aziendale
- una prospettiva logica incentrata sullo schema del data warehouse
- una prospettiva fisica che rappresenta il livello di trasporto fisico dei dati.

Queste prospettive corrispondono ai tre livelli tradizionali di data warehousing, vale a dire fonti, data warehouse e clienti.
La metodologia associa a ciascuna prospettiva una vista di metadati corrispondente, chiamata misura della qualità. Dal punto di vista della qualità dei dati, quattro fasi principali caratterizzano tale metodologia:

1. Definizione
2. Valutazione
3. Analisi
4. Miglioramento.

Uno dei principali contributi forniti da questa metodologia è la classificazione di dati e software nel contesto del data warehouse. Sono definite tre categorie di dati e metadati in questo contesto:

- Qualità di progettazione e amministrazione: la prima si riferisce alla capacità di un modello di rappresentare le informazioni in modo adeguato ed efficiente, mentre la seconda si riferisce al modo in cui il modello si evolve durante l'operazione di data warehouse.
- Qualità dell'implementazione del software: le dimensioni della qualità della norma ISO 9126 vengono presi in considerazione, poiché l'implementazione del software non è un'attività con caratteristiche specifiche del data warehouse.
- Qualità di utilizzo dei dati: si riferisce alle dimensioni che caratterizzano l'utilizzo e l'interrogazione dei dati contenuti nel data warehouse.

- **La metodologia TIQM (Total Information Quality Management)**

Questa metodologia è stata sviluppata per supportare i warehouse project ed assume la consolidazione di fonti di dati operativi in un database unico e integrato, utilizzato in tutti i tipi di aggregazioni eseguite per costruire il data warehouse (Batini e et al, 2009).

Questo consolidamento elimina gli errori e le diversità dei database di origine. La TIQM si concentra maggiormente sulle attività di gestione che sono responsabili dell'integrazione delle fonti di dati operativi, discutendo la strategia che deve essere seguita dall'organizzazione per compiere scelte tecniche efficaci. La metodologia fornisce una classificazione dettagliata di costi e benefici

La Figura 6 mostra le fasi della metodologia TIQM. Dal punto di vista manageriale del TIQM, ci sono tre fasi principali:

- Valutazione
- Miglioramento
- Gestione e monitoraggio del miglioramento.

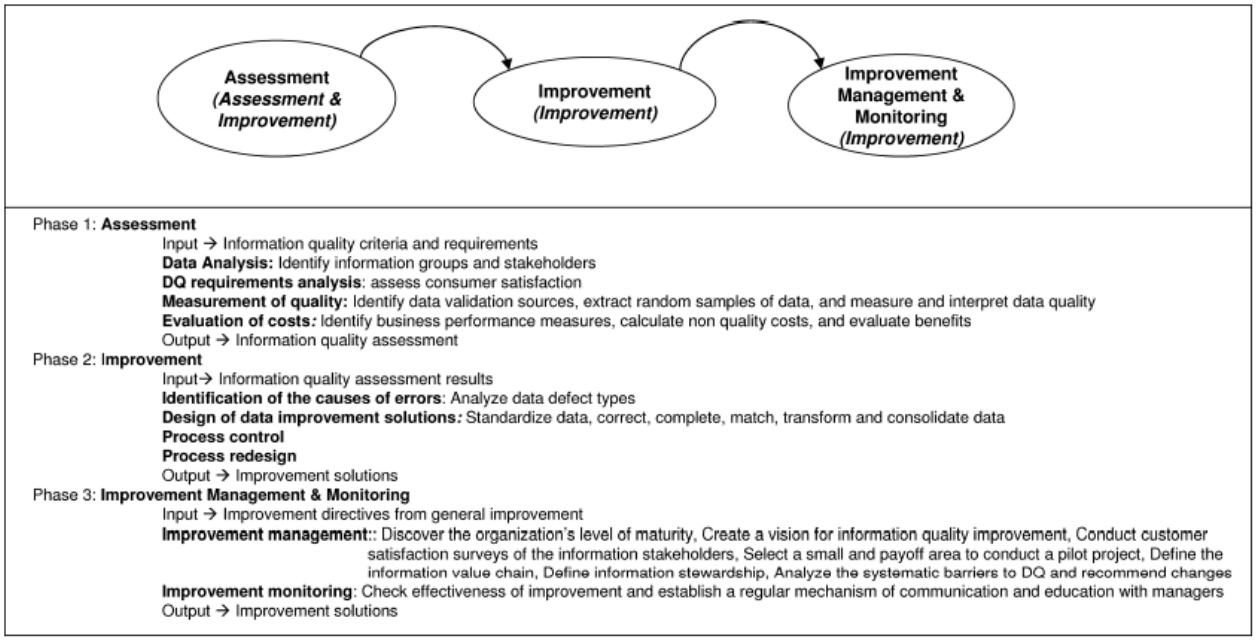

Figura 6 TIQM Fasi e processo (Batini e et al, 2009)

Uno dei preziosi contributi della metodologia è la definizione di quest'ultima fase, che fornisce le linee guida per gestire i cambiamenti nella struttura dell'organizzazione in base ai requisiti di gestione della qualità dei dati.

Inoltre, l'approccio economico introduce la valutazione costi-benefici per giustificare gli interventi sulla qualità dei dati. L'obiettivo non è solo il raggiungimento di dati più elevati a livello qualitativo, ma deve intraprendere azioni di miglioramento solo se realizzabili; quindi, solo se i benefici sono maggiori dei costi.

- **La metodologia AIMQ (Methodology for Information Quality Assessment).**

La metodologia AIMQ è l'unica metodologia di qualità dell'informazione focalizzata sul benchmarking (comparazione dei dati su più fronti) che è una tecnica oggettiva e indipendente dal dominio per la valutazione della qualità (Batini e et al, 2009).

La base della metodologia AIMQ è una tabella 2x2, denominata modello PSP/IQ (vedi Figura 7) che classifica le dimensioni della qualità in base alla loro importanza dal punto di vista dell'utente e del manager.

	Conforms to specifications	*Meets or exceeds consumer expectations*
Product Quality	Sound Information	Useful Information
Service Quality	Dependable Information	Usable Information

Figura 7 PSP/IQ model (Batini e et al, 2009)

Gli assi della tabella sono la conformità alle specifiche e la conformità alle aspettative degli utenti.

Di conseguenza, si distinguono quattro classi di dimensioni (“sound”, affidabile, utile e utilizzabile) e le dimensioni di qualità identificate sono classificati in queste classi.

Il “benchmarking” dovrebbe classificare le informazioni all'interno di ciascuna classe.
Il modello PSP/IQ è un input della metodologia AIMQ le cui fasi sono riassunte nella Figura qui sotto (Figura 8).

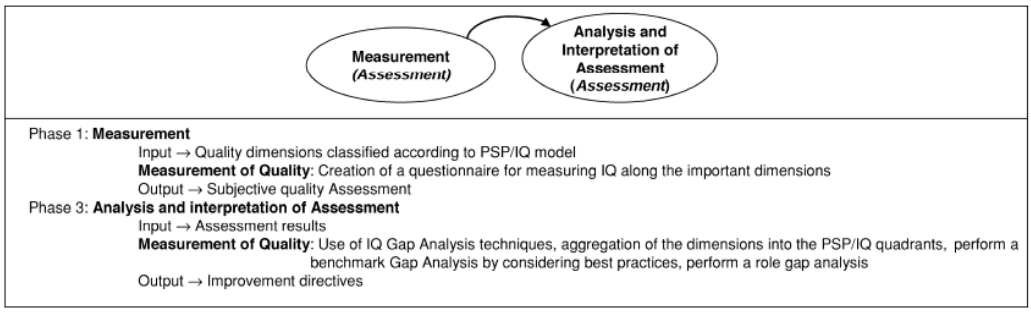

Figura 8 AIMQ Fasi (Batini e et al, 2009)

I lavori scientifici che descrivono l’AIMQ si concentrano principalmente sulle attività di valutazione, mentre le linee guida, le tecniche e gli strumenti per le attività di miglioramento non sono fornite.
Le tecniche di Gap Analysis sono consigliate come approccio standard per condurre benchmarking e interpretare i risultati.
In particolare, per questo approccio vengono suggerite due tecniche di “Gap Analysis”:

1. Information Quality Lacune nel benchmark
2. lacune nel ruolo della qualità delle informazioni.

Il primo confronta i valori della qualità di un'organizzazione con quelli delle organizzazioni di best practice. Quest'ultimo confronta le valutazioni della qualità delle informazioni fornite dai diversi ruoli organizzativi, ovvero il professionista IS e l'utente dell'informazione. L “IQ Role Gaps” ricerca le discrepanze tra le valutazioni fornite dai diversi ruoli come indicazione di potenziali problemi di qualità.

5. Conclusione

In questo lavoro sono stata introdotte le nozioni generali di base di dati e sono state discusse le funzionalità e le proprietà che caratterizzano la gestione dei dati.

La “Data Governance” è stato definita ed in particolar modo, sono state discusse le potenzialità e le caratteristiche. Infatti, grazie ad essa i dati vengono valorizzati e sfruttati per aumentare l’efficienza e la produttività.

Questo tipo di sistema è la tecnologia più adatta per sviluppare applicazioni che usano in modo ricorrente dati persistenti, in contesti in cui sia necessario accedere alle stesse informazioni da parte di più applicazioni, e dove è prevista un’evoluzione nel tempo delle esigenze di archiviazione e gestione di informazioni.

Le metodologie di miglioramento della qualità dei dati sono molteplici e devono essere implementate all’interno dei processi informatici in relazione alle caratteristiche del progetto.

DWQ, TIQM, AIQM sono solo alcune delle più comuni tecniche per il miglioramento e per la valutazione dei dati.

In definitiva la qualità dei dati, così come la loro gestione prevedono dei processi complessi che devono essere attentamente analizzati anche in relazione ai costi di applicazione.

Così come introdotto in precedenza, il costo della qualità dei dati (somma del costo delle attività di valutazione e costo associato alla scarsa qualità dei dati) può assumere un aspetto importante nel controllo di un progetto di qualsiasi contesto.

Bibliografia

Abdul Aziz, Azwa, Md Yazid Mohd Saman, and Mohd Po. «Using metadata analysis and base analysis techniques in DQ framework for DW.» 210: 608-613.

Albano, Antonio, Giorgio Ghelli, e Renzo Orsini. *FONDAMENTI DI BASI DI DATI.* Universita di Pisa, Universita di Venezia, 2021.

Batini, Barone, Mastrella, Maurino, e Ruffini. «QUALITY, A FRAMEWORK AND A METHODOLOGY FOR DATA.» s.d.

Batini, Carlo, e et al. «Methodologies for data quality assessment and improvement.» In *ACM computing surveys (CSUR) 41.3*, 1-52. 2009.

Batini, Carlo, e Monica Scannapieco. *Data and Information Quality - Dimensions, Principles and Technique.* Springer, 2016.

Bininda-Emonds, e Olaf RP et al. «Garbage in, garbage out.» *Springer, Dordrecht.*, 2004: 267-280.

Casadei, G., e A.G.B. Teolis . «LA GESTIONE DEI PROGETTI INFORMATICI.» 2013.

Chianese, A., A. Picariello, V. Moscato, e L. Sansone. *Basi di dati per la gestione dell'informazione.* 2007.

Chianese, A., V. Moscato, A. Picariello, e L. Sansone. *Sistemi di Basi di Dati e applicazioni.* 2015.

J., Bair. «Practical Data Quality: Sophistication Levels.» 2006.

Janssen, e Marijn et al. «Data governance: Organizing data for trustworthy Artificial Intelligence.» *Government Information Quarterly 37.3*, 2020.

JEUSFELD, M., QUIX, C., AND JARKE. «Design and analysis of quality information for data warehouses.» 1998.

Keller, KL, e R Staelin. «Effects of quality and quantity of information on decision effectiveness. .» *Journal of Consumer Research*, 1987: 200-213.

Khatri, V., e C.V, Brown. «Designing data governance. Communications of the ACM.» 2010: 148-152.

Lucarelli, Piero. s.d. http://www.pierolucarelli.it/tutorials%20office/access2000/database.htm.

Marinos. «We're Not Doing What? The Top 10 Corporate Oversights in Data Governance. DM Review, .» 2004.

Marinos,, G. «Data Quality: The Hidden Assumption Behind COSO.» 2004: 12.

Di Nunzio, Universita di Padova. «http://www.dei.unipd.it/~dinunzio/fdi-2014-2015/04_dati_informazioni.pdf.» 2014.

Olson, J. «Data Quality: The Accuracy Dimension.» *Morgan Kaufmann Publishers, USA*, 2003.

Otto, B. «Organizing data governance: Findings from the telecommunications industry and consequences for large service providers.» 2011: 45–66.

Otto, B., e K. Weber. . «Data governance." Daten-und Informationsqualität.".» Vieweg+ Teubner, 2011.

Technical committee ISO/TC 258 Project, programme and portfolio management. ISO 21500: Project Management. s.d.