UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

FACOLTÀ DI SCIENZE MATEMATICHE FISICHE E NATURALI

CORSO DI LAUREA IN INFORMATICA E TECNOLOGIE PER LA PRODUZIONE DEL SOFTWARE

TESI DI LAUREA

IN

MODELLI E METODI PER LA QUALITÀ DEL SOFTWARE

ANALISI DELLA QUALITA' DEI DATI: PROGETTAZIONE ED ESTENSIONE DEI TOOL PREESISTENTI

Relatori:

Chiar.ma Prof.ssa Maria Teresa Baldassarre

Laureanda: **Alessandra Cicciarelli**

Anno Accademico 2022/2023

Alla mia famiglia.

E a te, che ci guardi dall'alto.

ABSTRACT

Questa tesi si propone di analizzare il mercato dei sistemi software per la qualità dei dati, mettendo in luce le loro forze e debolezze al fine di migliorarli. Sebbene l'obiettivo iniziale fosse la selezione del software più idoneo per la gestione della qualità dei dati, durante lo studio si è manifestato un significativo cambiamento di approccio. La sfida di estendere efficacemente la soluzione selezionata ha orientato la ricerca in una nuova direzione.

Il principale scopo di questa tesi è ora fornire un modello guida per la scelta del miglior strumento di gestione della qualità dei dati in contesti aziendali. Inoltre, si intende sviluppare un prototipo di componente software per la gestione del controllo di qualità dei dati, dimostrando la fattibilità di un'implementazione da zero. L'obiettivo primario è contribuire all'ottimizzazione della gestione della qualità dei dati nelle organizzazioni, offrendo sia un modello guida per la selezione di strumenti appropriati, sia un esempio concreto di implementazione di controlli di qualità dei dati attraverso il prototipo sviluppato. Tale approccio permetterà alle organizzazioni di effettuare scelte più efficaci e migliorare la qualità dei propri dati, un elemento fondamentale per il successo aziendale nell'era dell'informazione.

INDICE

1. INTRODUZIONE

Attualmente la mole di dati gestiti dai sistemi informatici è in netto aumento. I dati costituiscono una risorsa fondamentale per tutte le aziende che li utilizzano nelle loro attività, che siano operative, decisionali, strategiche o di analisi.

Possedere dei dati errati o non conformi agli standard impedisce lo scambio di informazioni, tra sistemi informatici e utenti e tra sistemi informatici stessi; potrebbe in aggiunta portare a errori, analisi inaccurate, danni a livello economico ma anche d'immagine.

È pertanto importante integrare all'interno delle aziende delle figure preposte all'analisi e al miglioramento continuo della qualità dei dati, assicurando quindi accuratezza e affidabilità degli stessi a partire dalla loro raccolta dei dati fino allo scambio di informazioni.

Per comprendere al meglio come l'accuratezza e l'affidabilità dei dati possano influenzare l'efficacia delle attività aziendali e le decisioni strategiche, è fondamentale esplorare il concetto di Data Quality.

1.1 Data Quality: definizione

La qualità dei dati si riferisce alla misura in cui i dati sono accurati, completi, affidabili, tempestivi, consistenti e pertinenti per soddisfare le esigenze specifiche di un'organizzazione.

Rappresenta la capacità dei dati di svolgere il loro ruolo in modo efficace nelle diverse attività aziendali. **[1]**

La definizione di qualità dei dati può variare leggermente in base al contesto e agli obiettivi aziendali, ma ci sono alcune caratteristiche chiave:

- **Accuratezza**: I dati devono essere esatti e veritieri. L'accuratezza è fondamentale perché dati errati possono portare a decisioni sbagliate o analisi fuorvianti.

- **Completezza**: I dati devono essere completi, ovvero non devono mancare informazioni essenziali. Dati incompleti possono generare ambiguità e limitare la loro utilità.
- **Affidabilità**: I dati devono provenire da fonti attendibili e essere raccolti in modo coerente nel tempo. L'affidabilità è cruciale per la coerenza delle informazioni.
- **Tempestività**: I dati devono essere costantemente aggiornati e disponibili. L'obsolescenza dei dati può ridurne la loro utilità.
- **Consistenza**: I dati devono essere coerenti e non devono presentarsi conflitti tra gli stessi valori di dati in sistemi differenti.
- **Pertinenza:** I dati devono essere pertinenti alle esigenze e agli obiettivi dell'organizzazione. La raccolta di dati non pertinenti può rappresentare uno spreco di risorse.

L'importanza della qualità dei dati è evidente in vari contesti, tra cui:

- Business Intelligence e Analisi: Decisioni aziendali basate su analisi dei dati richiedono dati di alta qualità per essere accurate e significative.
- Gestione delle Relazioni con i Clienti (CRM): Dati di bassa qualità possono portare a una gestione inadeguata delle relazioni con i clienti, danneggiando la reputazione dell'azienda.
- Conformità normativa: Molte normative richiedono dati precisi e affidabili, e la mancanza di qualità dei dati può comportare multe e sanzioni.
- Automazione dei Processi Aziendali: La qualità dei dati è essenziale per l'automazione dei processi aziendali, poiché i sistemi automatizzati si basano su dati accurati.

Pertanto, garantire la qualità dei dati è un obiettivo cruciale per qualsiasi organizzazione. Ciò richiede processi di raccolta, gestione e controllo dei dati adeguati, oltre all'adozione di strumenti e tecnologie apposite per il miglioramento continuo della qualità dei dati.

Nel paragrafo successivo, porremo l'attenzione sugli standard internazionali che regolano le caratteristiche della qualità dei dati e le misurazioni delle stesse.

1.2 Data Quality: Standard ISO/IEC

L'Organizzazione internazionale per la standardizzazione (ISO) e la Commissione elettrotecnica internazionale (IEC) costituiscono il sistema specializzato per la standardizzazione mondiale.

La serie ISO/IEC 25000, denominata SQuaRE "Systems and Software Quality Requirements and Evaluation", è stata sviluppata dall'ISO/IEC JTC1 SC7, un comitato tecnico istituito per gli standard relativi ai prodotti informatici.

La SQuaRE è suddivisa in diverse sezioni, ciascuna delle quali si focalizza su un aspetto della qualità dei dati e del software. Queste le sezioni principali: **[2]**

- **ISO/IEC 2500n**: gestione della qualità (25000);
- **ISO/IEC 2501n**: modelli di qualità del software (25010), servizi (TS 25011), dati (25012), qualità in uso (25010 e TS 25011);
- **ISO/IEC 2502n**: misurazione della qualità come generalità (25020), elementi di misura (25021), qualità in uso (25022), software (25023), dati (25024) e servizi (CD 25025);
- **ISO/IEC 2503n**: requisiti di qualità (25030);
- **ISO/IEC 2504n**: processo di valutazione (25040), guida per sviluppatori (25041);
- **ISO/IEC 25050-99**: estensioni agli standard relative ai requisiti RUSP "Ready for Use Software Product", come le App, e istruzioni al testing (25051), framework per l'usabilità TR 25060), contesto d'uso (25063), report bisogni degli utenti (25064), formati sull'usabilità, report (DIS 25065), report di valutazione (25066).

Per quanto riguarda lo studio effettuato in questa tesi, relativo alla qualità dei dati, risulta necessario soffermarsi sulle norme ISO/IEC 25012 e 25024.

1.2.1 ISO/IEC 25012

Lo standard ISO/IEC 25012 fornisce un modello generale di qualità dei dati da utilizzare per organizzare i dati in maniera strutturata all'interno di un sistema informatico. **[5]**

Sono state individuate quindici caratteristiche della qualità dei dati suddivise secondo due punti di vista: la qualità intrinseca dei dati e la qualità dei dati dipendente dal sistema. (

Tabella 1)

Qualità Intrinseca dei Dati

Fa riferimento al grado con cui le caratteristiche di qualità dei dati hanno il potenziale intrinseco per soddisfare le esigenze esplicite e implicite. Da un punto di vista intrinseco, la qualità dei dati è legata ai dati stessi e comprende:

- *Dominio dei dati ed eventuali restrizioni*: questo aspetto si riferisce ai valori presenti nel dominio dei dati e alle eventuali restrizioni, come le regole aziendali, che definiscono la qualità richiesta per una specifica caratteristica in un'applicazione.
- *Relazioni tra i valori dei dati (consistenza e coerenza):* La coerenza tra i valori dei dati è di fondamentale importanza. Deve essere garantito che i dati non presentino conflitti o discrepanze quando utilizzati in diversi contesti.
- *Metadati:* I metadati forniscono informazioni essenziali sulla struttura e sul significato dei dati. Questi giocano un ruolo cruciale nell'assicurare la comprensione e l'uso corretto dei dati.

Qualità dei Dati Dipendente dal Sistema

Fa riferimento al grado con cui la qualità dei dati è raggiunta e mantenuta all'interno di un sistema informatico. Da questo punto di

vista, la qualità dei dati è influenzata dal dominio tecnologico in cui i dati vengono utilizzati, ed è raggiunta attraverso le capacità dei componenti dei sistemi informatici, tra cui:

- *Dispositivi hardware:* svolgono un ruolo chiave nel raggiungimento della qualità dei dati, ad esempio assicurando che i dati siano sempre disponibili o che venga raggiunta la precisione richiesta.
- *Sistema software:* software di base del sistema informatico, ad esempio il software di backup, essenziale per garantire la possibilità di recuperare i dati in caso di eventi critici.
- *Altri Software:* ad esempio strumenti di migrazione dati che sono fondamentali per garantire la portabilità dei dati tra diversi sistemi e contesti.

Questo approccio distinto tra qualità intrinseca dei dati e qualità dei dati dipendente dal sistema sottolinea l'importanza della qualità dei dati in ogni fase del loro ciclo di vita e nell'ambito delle attività aziendali.

Caratteristica	Inerente	Dipendente dal sistema
ACCURATEZZA	X	
COMPLETEZZA	X	
CONSISTENZA	X	
CREDIBILITA'	X	
ATTUALITA'	X	
ACCESSIBILITA'	X	X
CONFORMITA'	X	X
CONFIDENZIALITA'	X	X
EFFICIENZA	X	X

PRECISIONE	**X**	**X**
TRACCIABILITA'	**X**	**X**
COMPRENSIBILITA'	**X**	**X**
DISPONIBILITA'		**X**
PORTABILITA'		**X**
RECUPERABILITA'		**X**

Tabella 1: caratteristiche di qualità inerenti e dipendenti dal sistema.

1.2.2 ISO/IEC 25024

Lo standard ISO/IEC 25024 definisce le misure per la qualità dei dati che permettono di quantificare la qualità dei dati in base alle caratteristiche definite nell' ISO/IEC 25012.

Le caratteristiche di qualità possono essere quantificate applicando delle funzioni di misurazione, queste generalmente normalizzano i valori in un range che permette di verificare quanto i requisiti di qualità sono soddisfatti. [6]

Questo approccio permette alle aziende di valutare la qualità dei loro dati in maniera più completa e dettagliata, fornendo loro un set di misure di qualità dei dati che possono essere adattate alle loro esigenze specifiche e al contesto di utilizzo.

2. OBIETTIVO E SCOPO DELLA TESI

L’obiettivo di questa tesi è quello di analizzare il mercato dei sistemi software per la qualità dei dati, andando a evidenziare i punti di forza e di debolezza delle soluzioni e migliorarle.

L'idea iniziale era quella di selezionare, a partire dai sistemi valutati, il software più adatto e conforme alla gestione della qualità dei dati e di utilizzarlo come punto di partenza per un sistema che rispettasse le caratteristiche di qualità sopra descritte e che permettesse di soddisfare le esigenze degli utenti.

Tuttavia, durante il corso dello studio, si è verificato un cambiamento significativo nell'approccio. La sfida di estendere efficacemente la soluzione selezionata ha portato a una nuova direzione di ricerca. Di conseguenza, il principale scopo di questa tesi è stato adattato per rispondere a questa nuova prospettiva.

Lo scopo principale della tesi pertanto consiste nel fornire un modello di guida per la selezione dello strumento più adatto per la gestione della qualità dei dati in contesti aziendali.

In aggiunta, questa tesi si propone di sviluppare un prototipo di componente software per la gestione di un controllo di qualità dei dati, al fine di proporre nel modello di guida anche l’implementazione di un sistema da zero, dimostrando la sua fattibilità.

L'obiettivo principale è quindi contribuire all'ottimizzazione della gestione della qualità dei dati nelle organizzazioni, fornendo sia un modello di guida per la scelta di strumenti appropriati sia un esempio pratico di implementazione di controlli di qualità dei dati attraverso lo sviluppo del prototipo. Questo approccio consentirà alle organizzazioni di effettuare scelte efficaci e di migliorare la qualità dei loro dati, un elemento cruciale per il successo aziendale nell'era dell'informazione.

3. BACKGROUND

3.1 Contesto

La Data Quality (DQ) è definita come il soddisfacimento del "FIT FOR USE" auspicato da coloro che utilizzano i dati per consentire operazioni aziendali efficienti e decisioni concise. Per raggiungere questo obiettivo è emerso un mercato per gli strumenti di DQ strettamente integrato con i mercati dell'integrazione dei dati o della gestione dei metadati.
Gli strumenti per la DQ, quindi, si concentrano sul supporto delle diverse
fasi del ciclo di vita della DQ, definendo, misurando, analizzando e migliorando la qualità del set di dati.

Inizialmente, gli strumenti di DQ si concentravano sul soddisfare le linee guida interne di conformità e sulla riduzione dei rischi mediante la definizione manuale di regole che i nuovi dati avrebbero dovuto rispettare. Attualmente questi strumenti stanno progredendo grazie alla continua crescita dell'automazione e dell'intelligenza artificiale.

Possiamo distinguere i vari strumenti in tre macrocategorie che ne identificano le funzionalità:

- Strumenti di **data profiling**: il profiling dei dati è il processo di analisi dei dati per identificare problemi di qualità, come valori mancanti, valori duplicati, valori fuori scala e altri errori. Gli strumenti di profilazione dei dati forniscono statistiche e visualizzazioni che aiutano a comprendere la struttura e la qualità dei dati.
- Strumenti di **data cleansing**: La pulizia dei dati è il processo di correzione e standardizzazione dei dati per eliminare errori, valori duplicati o ambigui. Gli strumenti di pulizia dei dati consentono di applicare regole di trasformazione e di rimuovere o correggere dati errati.
- Strumenti di **data quality management** (***DQM***): Il DQM è un approccio completo che incorpora profiling, pulizia, controllo della qualità e monitoraggio dei dati. Questi strumenti

forniscono una piattaforma completa per garantire la qualità continua dei dati.

Oltre alle soluzioni esistenti, gli strumenti di DQ spesso vengono creati come soluzioni personalizzate. In questo modo, si può ottenere un'integrazione senza intoppi con gli strumenti esistenti di gestione dei dati e il resto del panorama IT.

3.2 Analisi delle soluzioni esistenti

Per fornire un modello ben preciso delle linee guida da seguire nella scelta dello strumento da utilizzare per la gestione della qualità dei dati è bene partire dalle soluzioni esistenti. In questo paragrafo si forniscono i punti di forza e di debolezza delle soluzioni analizzate durante questo studio, in modo tale da poter effettuare una corretta valutazione in fase di scelta.

Il punto di partenza dell'analisi è il Quadrante Magico di Gartner (Figura 1), uno strumento di analisi e ricerca ampiamente riconosciuto nel settore tecnologico. Viene utilizzato per valutare e posizionare graficamente le aziende in un determinato mercato in base a due dimensioni principali: la loro capacità di esecuzione e la completezza della loro visione. **[9]**

3.2.1 Quadrante Magico di Gartner

Il Quadrante Magico di solito presenta quattro categorie principali in cui le aziende vengono posizionate: **[8]**

- Leader: Le aziende in questa categoria sono considerate tra le migliori nel mercato. Hanno dimostrato una forte capacità di esecuzione, offrendo prodotti o servizi affidabili e completi. Inoltre, hanno una visione chiara e innovativa del futuro del mercato nel lungo termine.
- Challengers: Le aziende in questa categoria hanno una forte capacità di esecuzione, ma la loro visione di mercato a lungo termine potrebbe non essere altrettanto chiara o avanzata come quella dei leader. Tuttavia, sono in grado di competere con successo nel mercato grazie alla loro forza operativa.

- Visionaries: Queste aziende hanno una visione innovativa e avanzata per il mercato, ma potrebbero avere delle sfide nella capacità di esecuzione. Possono offrire prodotti o servizi unici e all'avanguardia, ma potrebbero dover lavorare sulla loro maturità operativa.
- Niche Players: Le aziende in questa categoria potrebbero avere una buona capacità di esecuzione in un piccolo segmento di mercato specifico ma potrebbero non essere così innovative o ampie nella loro visione. Sono focalizzate su mercati di nicchia o segmenti specializzati.

Le differenze tra queste quattro categorie si basano principalmente su due fattori principali: la capacità di esecuzione e la completezza della visione.

- Capacità di Esecuzione: Questo aspetto riflette la capacità dell'azienda di consegnare prodotti o servizi in modo affidabile, soddisfacendo le esigenze dei clienti, rispettando gli accordi contrattuali e mantenendo una solida presenza di mercato. Le aziende posizionate come leader o sfidanti di solito hanno una forte capacità di esecuzione.
- Completezza della Visione: Questo aspetto riflette la capacità dell'azienda di comprendere e anticipare le tendenze del mercato, oltre a sviluppare una visione chiara per il futuro. Le aziende posizionate come visionarie o leader di solito hanno una visione più ampia e innovativa.

Figura 1: Quadrante Magico di Gartner per le soluzioni di Data Quality

3.2.2 Valutazione dei Leaders

Per lo studio sono stati considerati i fornitori individuati come Leader da Gartner, in modo tale da potersi focalizzare sui fornitori dei servizi più completi.

Di seguito la valutazione dei punti di forza e dei punti di debolezza dei sei fornitori leader: **[3]**

FORNITORE	PUNTI DI FORZA	PUNTI DI DEBOLEZZA
Informatica	• Utilizzo dell'intelligenza artificiale per individuare anomalie nei dati e convertirle in regole riutilizzabili.	• Costi aggiuntivi per la migrazione al cloud • Interfaccia utente e interfaccia report migliorabili

	• Associazione intelligente delle regole e automazione dei flussi di lavoro di correzione • Piattaforma cloud con gestione della migrazione semplificata • Presenza di un developer tool	
IBM	• Ampia suite di strumenti disponibile, tra cui strumenti di profiling, cleansing, integrazione e monitoraggio dei dati in modo da gestire i dati in tutto il loro ciclo di vita • Supporto clienti ben gestito, con l'integrazione dell'intelligenza artificiale attraverso IBM Watson • Utilizzo dell'intelligenza artificiale per l'automazione delle attività di gestione dei dati,	• Costi elevati • Complessità d'utilizzo, dovuta proprio alle diverse funzioni fornite, che potrebbe renderlo meno accessibile per utenti meno esperti o per piccole aziende • Scarsa chiarezza sulle licenze necessarie, spesso delle funzionalità richiedono licenze separate, quindi costi aggiuntivi

	come il riconoscimento dei pattern, l'auto-correzione • Presenza di un toolkit per gli sviluppatori con diversi tutorial disponibili	
SAP	• Ampia suite di funzionalità per la gestione della qualità dei dati, inclusa la registrazione degli accessi ai dati fondamentale per la verificabilità • Capacità di gestione dei dati Master MDM • Interfaccia utente intuitiva • Scalabilità elevata	• Costi elevati • Difficoltà nella personalizzazione • Complessità nell'utilizzo • Integrazione limitata con prodotti non SAP
Talend	• Capacità di data profiling, standardizzazione, pulizia, matching e merging dei dati • Integrazione flessibile con altre applicazioni • Open source • Presenza di una community	• Supporto clienti migliorabile • Monitoraggio delle analisi e reporting non sufficiente per vari casi d'uso • Documentazione poco completa • Necessità di team esperti per la personalizzazione

	• Presenza di un kit (TCK) per lo sviluppo di nuovi componenti, con la possibilità di integrarlo in un IDE attraverso un plugin o di utilizzare un'interfaccia web per la creazione del nuovo componente	
SAS	• Ampia suite di prodotti per la data quality: data management, data preparation, data governance • Migrazione al servizio cloud con supporto open source • Funzionalità di autoapprendimento basate sul machine learning • Ottimo supporto ai clienti	• Difficoltà nell'installazione e aggiornamento • Interfaccia utente migliorabile • Difficoltà di integrazione con sistemi non SAS • Mancanza di automatizzazione in alcune fasi della gestione della qualità dei dati
Ataccama	• Impiego dell'intelligenza artificiale in diverse funzionalità quali	• Interfaccia per la visualizzazione dei dati migliorabile • Difficoltà nella personalizzazione

	rilevamento di errori e suggerimenti automatici • Integrazione delle varie funzionalità all'interno di un'unica soluzione, quali profiling, gestione dei dati master e gestione dei metadati • Correzione dei dati tempestiva	della soluzione • Documentazione poco chiara e completa

3.2.3 Valutazione dei Challengers

FORNITORE	PUNTI DI FORZA	PUNTI DI DEBOLEZZA
Experian	• Diverse funzionalità di convalida di dati inclusi indirizzi, email, telefoni e dati geografici • Facilità d'uso e d'implementazione	• Risulta arretrato nell'innovazione rispetto ad altri concorrenti, infatti non sono fornite soluzioni di apprendimento automatico • Migliorabile dal punto di vista dell'analisi e della standardizzazione dei dati • Costoso rispetto al servizio fornito
Innovative Systems	• Inclusione dell'intelligenza	• Risulta essere poco prestante nei

	artificiale nell'analisi, standardizzazione e pulizia dei dati • Supporto clienti ben gestito • Costi non elevati	processi batch che utilizzano molta memoria e risultano lenti con file di grandi dimensioni • Scalabilità limitata
Redpoint	• Facilità nell'installazione, aggiornamento e utilizzo della soluzione • Risulta essere prestante anche con molti record da gestire • Permette di definire regole di qualità dei dati in base alle esigenze	• Complessità nella personalizzazione • Difficoltà nell'integrazione con altre soluzioni

3.2.4 Valutazione dei Visionaires

FORNITORE	PUNTI DI FORZA	PUNTI DI DEBOLEZZA
Collibra	• Automazione delle attività legate alla gestione di qualità dei dati • Supporto di dati provenienti da diverse sorgenti • Interfaccia intuitiva per la definizione delle regole per l'analisi	• Complessità nell'integrazione con altri sistemi • Scarsa adozione degli standard di qualità

	di qualità	
Syniti	• Integrazione con sistemi SAP • Offre funzionalità adatte a diversi casi d'uso • Automatizzazione delle fasi di pulizia e gestione dei dati	• Minori capacità di data profiling rispetto ai concorrenti • Difficoltà con la gestione di risorse esterne
Precisely	• Buona gestione delle funzionalità principali della qualità dei dati quali standardizzazione e pulizia, ma anche matching e merging dei dati • Funzionalità di convalida di indirizzi e dati geografici	• Difficoltà nell'installazione, aggiornamento e integrazione con altri sistemi • Mancanza di una documentazione chiara e completa
TIBCO Software	• Analisi avanzate dei dati in tempo reale • Elevata scalabiltà • Gestione dell'integrazione e dell'analisi di qualità attraverso apprendimento automatico	• Complessità nell'utilizzo • Difficoltà nella personalizzazione della soluzione, con un ampio impiego di risorse

3.2.5 Valutazione dei Niche Players

FORNITORE	PUNTI DI FORZA	PUNTI DI DEBOLEZZA
MIOSoft	• Facilità	• Non dispone di

	nell'installazione, utilizzo e aggiornamento • Scalabilità e affidabilità alte	una community di supporto per i clienti • Documentazione poco leggibile • Funzionalità per l'analisi della qualità non conformi agli standard
Datactics	• Conformità alle norme ISO nell'analisi della qualità dei dati • Facilità nell'utilizzo	• Complessità nell'integrazione con altri sistemi • Difficoltà nella personalizzazione della soluzione fornita
Melissa	• Ampia suite per la validazione e standardizzazione dei dati • Supporto ai clienti ben gestito • Documentazione chiara e leggibile	• Scarsa automatizzazione dei processi di gestione dei dati e della data preparation

4. TALEND: ANALISI DEL TOOL

4.1 Motivazioni della scelta

Dopo un'attenta analisi delle diverse opzioni disponibili sul mercato, è risultato opportuno adottare la soluzione fornita da Talend per lo studio in questione. Questa decisione è stata presa a seguito di una valutazione dettagliata delle caratteristiche e delle capacità offerte da Talend, che lo rendono la scelta ideale per gli obiettivi della tesi, in particolare per quanto riguarda il raggiungimento degli standard di qualità dei dati previsti dalle normative ISO.

Talend è una piattaforma di integrazione dati open source che offre diversi software e servizi dedicati all'ambito dell'integrazione dati. Le sue funzionalità includono la gestione dei dati, l'integrazione delle applicazioni aziendali e il controllo e il miglioramento della qualità dei dati. Ciò significa che Talend fornisce una suite completa di strumenti per aiutare le organizzazioni a gestire, analizzare e ottimizzare i loro dati in modo efficiente e conforme agli standard di qualità richiesti.

In particolare, le caratteristiche che hanno portato alla scelta di Talend includono la flessibilità, la capacità di gestire volumi elevati di dati, la scalabilità per adattarsi alle esigenze in continua evoluzione e la sua natura open source, che consente una maggiore personalizzazione e flessibilità nell'implementazione delle soluzioni di integrazione dati.

4.2 Talend Data Fabric

Talend fornisce un’unica piattaforma modulare per integrazione, integrità e governance dei dati: Talend Data Fabric.

Un data fabric è un unico ambiente con un’architettura unificata, costituito da servizi che puntano a massimizzare l’efficienza della gestione dei dati. [7]

Talend Data Fabric rappresenta una soluzione completa che soddisfa appieno le esigenze delle moderne imprese orientate ai

dati, offrendo un ambiente unificato con un'architettura nativa che consente di adattarsi in modo rapido ai cambiamenti, garantendo al contempo l'integrità dei dati. I tratti distintivi di questa soluzione si traducono nell'abilità di fornire dati affidabili, integri, puliti e non compromessi.

Ambiente Unificato: Talend mette a disposizione un ambiente unificato in grado di soddisfare tutte le necessità relative alla gestione dei dati, facilitando la trasformazione dei dati grezzi in dati affidabili. Talend Data Fabric elimina la necessità di utilizzare diverse soluzioni, contratti e sistemi di supporto per l'integrazione dei dati. Copre l'intero ciclo, dalla scoperta e l'ingestione dei dati alla loro integrazione proveniente da diverse fonti, dalla pulizia dei dati all'assicurazione dell'integrità dei dati e all'analisi e condivisione di dati affidabili con tutte le parti interessate.

Generazione di Codice Nativo: Talend è in grado di generare codice nativo ottimizzato (in Java/Spark/SQL) durante la costruzione delle pipeline di dati, sfruttando le principali piattaforme come AWS, Azure o Snowflake. Questo, abbinato ai più di mille connettori e componenti di Talend per le applicazioni e gli ambienti più diffusi, semplifica notevolmente la scrittura del codice e la creazione delle pipeline.

On-Premise o in Cloud: Talend Data Fabric è progettato nativamente per operare sia in ambienti on-premise sia in cloud. Essa è in grado di ingegnare ed integrare dati da ambienti di back-office interni all'azienda, come Oracle e SAP, così come da ambienti cloud come AWS, Azure, Google Cloud o Snowflake.

Qualità e Governance dei Dati: Talend Data Fabric integra la qualità dei dati in tutte le fasi della gestione dati. Ciò include la scoperta e l'ingestione dei dati, l'utilizzo di Talend per la data stewardship e la definizione dei ruoli per la pulizia dei dati, nonché la tracciatura dell'origine dei dati per garantirne la conformità e l'integrità. La soluzione è progettata per consentire la condivisione di dati affidabili tramite una gestione dei dati in modalità self-service.

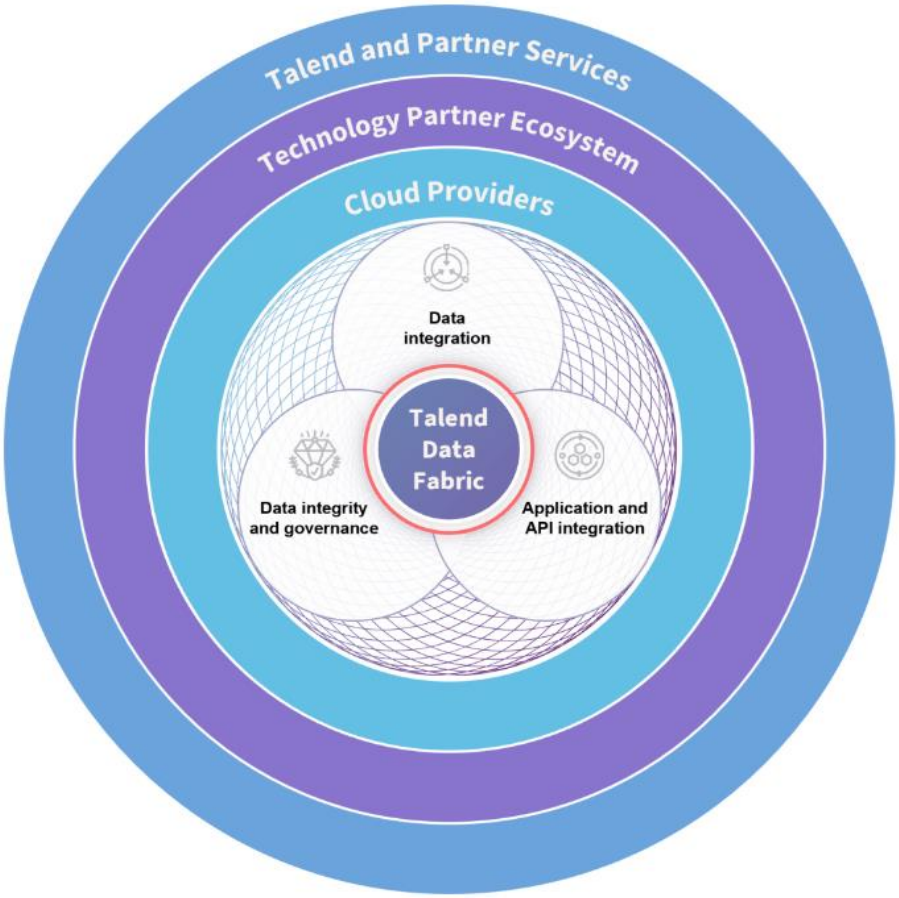


Figura 2: Talend Data Fabric

4.3 Talend Data Quality

Talend Data Quality è parte integrante di Talend Data Fabric, permette di eseguire profilazione dei dati, identificando rapidamente i problemi di qualità. Fornisce diverse funzioni per il miglioramento della qualità, tra le quali deduplicazione, convalida e standardizzazione. Permette anche di revisionare le analisi di qualità effettuate attraverso report e statistiche riepilogative.

Per lo studio, è stata scaricata la versione gratuita di Talend Data Quality in modo tale da poter avere una visione completa delle sue funzionalità, che verranno di seguito descritte.

4.3.1 Architettura Talend Data Quality

L'architettura di Talend Open Studio for Data Quality è composta da diversi blocchi funzionali (Figura 3):

- Il blocco Client include la piattaforma Talend Studio
- Il blocco Connection include le connessioni alle fonti di dati, database, big data e file delimitati.
- Il blocco Data Profiling riguarda la prospettiva per l'analisi dei dati, da qui l'utente può utilizzare pattern e indicatori predefiniti da Talend o customizzarli per analizzare i dati per cui sono state create delle connessioni
- Il Blocco Data Explorer riguarda la prospettiva che permette di visualizzare i risultati ottenuti dalle analisi effettuate sui dati.

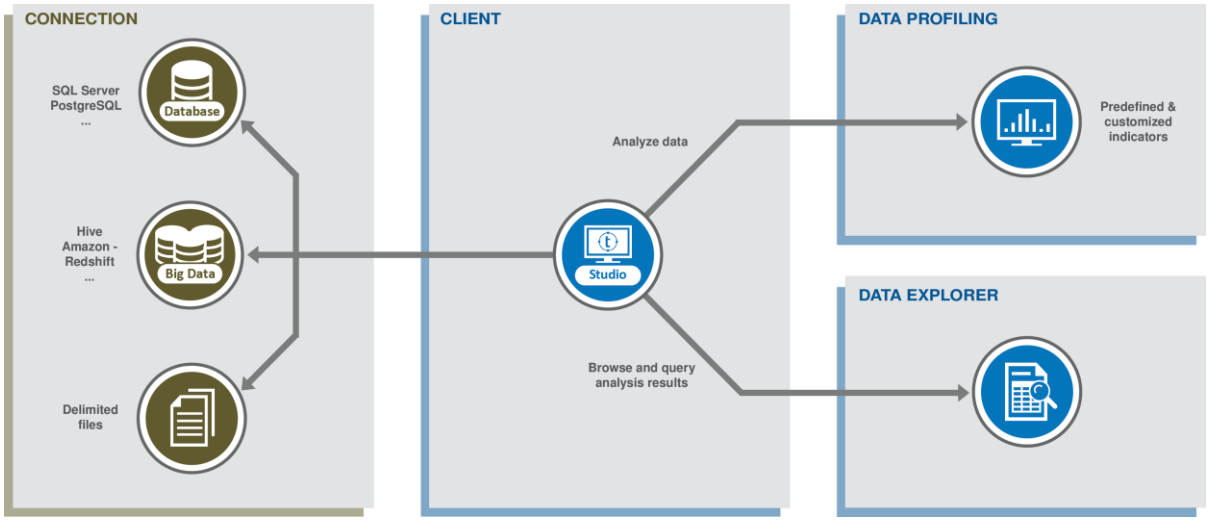


Figura 3: Architettura Talend for Data Quality

4.3.2 Data Profiling: Pattern e Indicatori

La prospettiva di Profiling include un data profiler da cui accedere alle analisi effettuate o crearne di nuove, una serie di librerie che comprendono Pattern e Indicatori e una sezione per la gestione dei metadati. Ci soffermeremo sui pattern (Figura 4) e indicatori (Figura 5), poiché a partire da questi sono definite le regole per l'analisi di qualità.

Pattern

I pattern sono set di stringhe che permettono di effettuare il matching con il contenuto delle colonne da analizzare, in Talend sono presenti due tipologie di pattern utilizzabili entrambi sia con l'analisi delle colonne che con l'analisi di insiemi di colonne:

- Regex: sono dei modelli predefiniti utilizzati per cercare e manipolare il testo all'interno degli elementi delle colonne da analizzare, è possibile utilizzare le espressioni regolari fornite da Talend o crearne di proprie.
- SQL pattern: sono pattern personalizzati utilizzati nelle query SQL, personalizzabili e implementabili da zero, solitamente questi contengono il simbolo percentuale (%).

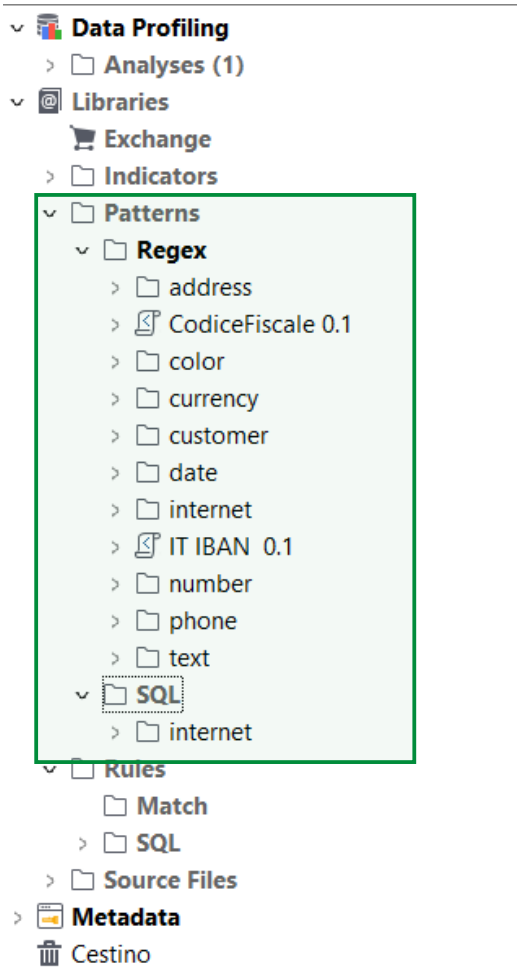


Figura 4: Pattern Talend Data Quality

Indicatori

Gli indicatori sono i risultati ottenuti attraverso l'implementazione di diversi modelli. Possono rappresentare i risultati della corrispondenza dei dati e diverse altre operazioni relative ai dati. La prospettiva Profiling di Talend Studio include due tipi di indicatori:

- Indicatori di sistema, un elenco di indicatori predefiniti raggruppati in diverse categorie (Simple statistics, Correlation, Tezt Statistics,..).
- Indicatori definiti dall'utente, un elenco di indicatori definiti dall'utente, questi vengono utilizzati solo nelle analisi delle colonne.

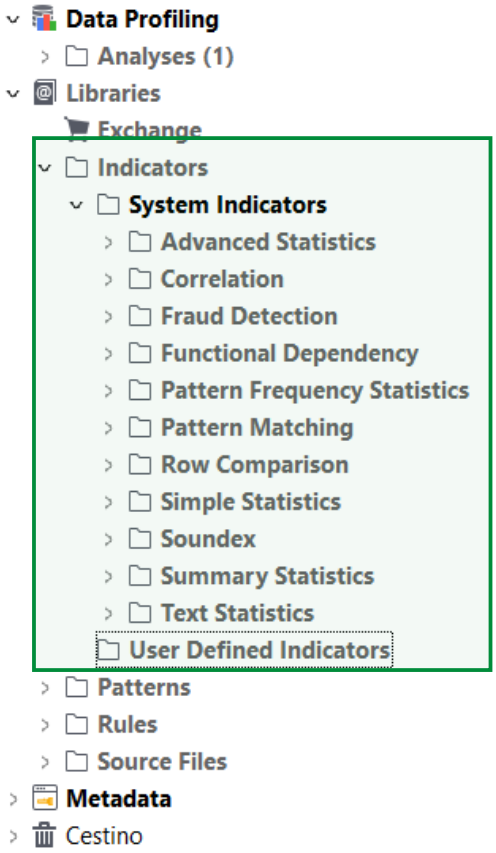


Figura 5: Indicatori Talend Data Quality

4.4 Talend Data Integration

Talend Data Integration fornisce una soluzione unificata che combina integrazione, trasformazione e mappatura rapida dei dati con controlli di qualità automatizzati. Permette di integrare dati da differenti sorgenti, di sviluppare e implementare facilmente pipeline di dati riutilizzabili per facilitare e velocizzare di gran lunga la scrittura tradizionale di codice.

Inoltre con Talend for Data Integration è possibile gestire sia i processi ETL (Extract/Transform/Load) che quelli ELT (Extract/Load/Transform).

4.4.1 Architettura Talend Data Integration

Come per Talend Data Quality, anche l'architettura di Talend Data Integration è composta da diversi blocchi funzionali che permettono di isolare le varie funzionalità. I diversi blocchi sono (Figura 6):

- Il blocco Client: include uno o più Talend Studio e browser Web che potrebbero trovarsi sulla stessa macchina o su macchine diverse.
- Il blocco Server comprende:
    - un server applicativo basato sul web, Talend Administration Center, che consente la gestione e l'amministrazione di tutti i progetti: i metadati di amministrazione (ad esempio account utente, diritti di accesso e autorizzazione del progetto) sono archiviati nel database di amministrazione, i dati degli elementi del progetto (ad esempio lavori, modelli di business e routine) sono archiviati nel server SVN o Git.
    - server utilizzati dalle applicazioni Web Talend, quali Talend Data Preparation, Talend Data Stewardship.
- Il blocco Repositories: include il server SVN o Git e il repository Nexus. Il server SVN o Git viene utilizzato per centralizzare tutti gli elementi del progetto come lavori e modelli di business condivisi tra diversi utenti finali ed è accessibile da Talend Studio per sviluppare elementi del progetto e da Talend Administration Center per pubblicare,

distribuire e monitorare gli elementi del progetto. Il repository Nexus viene utilizzato per archiviare aggiornamenti software disponibili per il download e lavori pubblicati da Talend Studio e pronti per essere distribuiti ed eseguiti.

- Il blocco Talend Execution Servers comprende uno o più server di esecuzione, distribuiti all'interno del sistema informativo dell'utente. I lavori Talend vengono distribuiti ai Job Server per essere eseguiti in un'ora, una data o un evento pianificati.
- Il blocco Database comprende i database di Amministrazione, Audit e Monitoraggio. Il database di amministrazione viene utilizzato per gestire gli account utente, i diritti di accesso e l'autorizzazione del progetto e così via. Il database Audit viene utilizzato per valutare diversi aspetti dei Job implementati nei progetti sviluppati in Talend Studio con l'obiettivo di fornire solidi fattori quantitativi e qualitativi per un supporto decisionale orientato al processo. Il database di monitoraggio viene utilizzato per monitorare le chiamate di servizio.

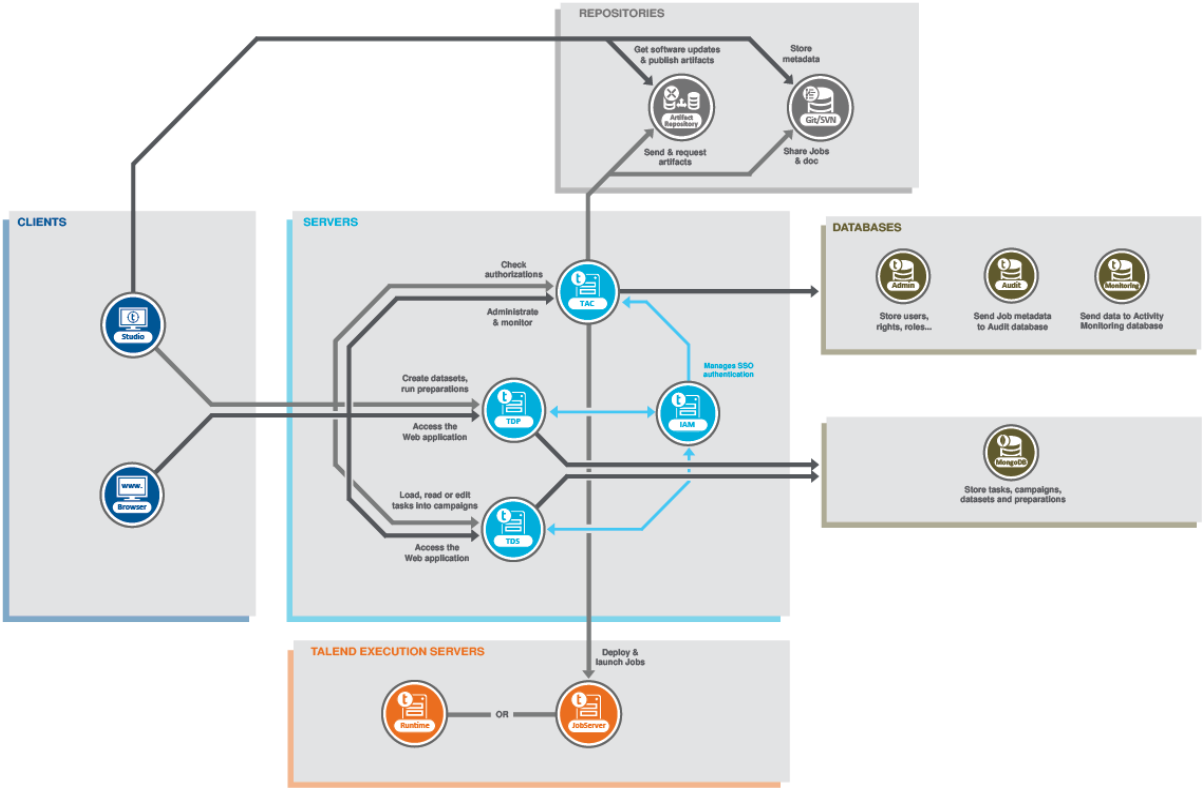


Figura 6: Architettura Talend Data Integration

4.5 Talend Component Kit

In seguito alla descrizione delle funzionalità principali delle due piattaforme di Talend rilevanti per lo studio, è bene evidenziare che nonostante entrambe le piattaforme permettessero la personalizzazione delle regole o dei pattern (in TDQ) e dei componenti (in TDI), l'obiettivo della tesi non era esattamente questo.

L'obiettivo infatti è la creazione di un nuovo componente, con una sua interfaccia e una sua funzionalità appropriata per dimostrare la personalizzazione effettiva della soluzione scelta. Pertanto è stato utilizzato il Talend Component Kit fornito da Talend come opzione per gli sviluppatori che intendono creare un nuovo componente nella Palette di Talend.

Talend Component Kit è un toolkit basato su Java e progettato per semplificare lo sviluppo di componenti a livello di: **[7]**

- Runtime: riguarda l'inserimento del codice del componente specifico in un lavoro o in una pipeline. Il framework aiuta a unificare il più

possibile il codice necessario per l'esecuzione in ambienti Data Integration e BEAM.

- Interfaccia grafica: il framework aiuta a unificare il codice richiesto per poter eseguire il rendering del componente in un browser (web) o nell'ambiente di sviluppo basato su Eclipse (SWT).

Il framework Talend Component Kit è composto da diversi strumenti progettati per facilitare il lavoro durante il processo di sviluppo dei componenti. Consente di sviluppare componenti che si adattano a entrambe le interfacce utente Web Java.

- Starter: genera lo scheletro del progetto di sviluppo utilizzando un'interfaccia user-friendly, è disponibile come strumento Web o come plug-in per l'IDE IntelliJ .
- API dei componenti: controlla tutte le classi disponibili per implementare i componenti.
- Strumenti di compilazione: il framework viene fornito con wrapper Maven e Gradle , che consentono di utilizzare sempre la versione di Maven o Gradle adatta all'ambiente e alla versione di sviluppo del componente.
- Strumenti di test: testa i componenti prima di integrarli nelle applicazioni Talend Studio o Cloud. Gli strumenti di test includono Talend Component Kit Web Tester, che consente di controllare l'interfaccia utente Web dei componenti sul computer locale.

Una volta generato il progetto, è possibile procedere con l'implementazione della logica e del layout dei componenti, iterando secondo le necessità. La fase di implementazione è costituita da alcuni passaggi principali che includono la definizione dei metadati della famiglia di cui farà parte il componente e dei componenti, l'implementazione della logica del componente di input, della logica di processore/uscita e della logica del componente autonomo, nonché la definizione del layout e delle configurazioni specifiche dei componenti.

Di conseguenza, è stata avviata la fase di sviluppo del nuovo componente. È stata condotta una fase preliminare di progettazione del componente. In questo contesto specifico, l'obiettivo principale era la creazione di un componente in grado di accettare in input un campo contenente nomi di città, confrontare

ogni elemento con le città presenti in un file scaricato dall'archivio di dati dell'Istat e restituire in output i risultati del controllo effettuato, identificando le città che richiedevano correzioni e fornendo al contempo possibili suggerimenti di correzione.

Il processo di progettazione di questo componente ha comportato diverse fasi ben definite. Innanzitutto, sono stati chiaramente delineati i requisiti e gli obiettivi del componente, tra cui la specifica del formato dei dati di input e del file per il confronto, oltre all'identificazione delle logiche di confronto e delle regole di correzione da applicare alle città non conformi. Successivamente, è stata sviluppata un'interfaccia utente per consentire agli utenti di configurare il componente in base alle loro specifiche esigenze.

Una delle sfide principali affrontate durante l'implementazione di questo componente è stata la gestione delle diverse eccezioni e condizioni speciali. Ad esempio, alcune città potrebbero presentare leggere discrepanze nella scrittura rispetto ai dati nel file ISTAT, oppure potrebbero esistere città con nomi molto simili, generando ambiguità durante il processo di correzione.

Inoltre, il componente doveva essere altamente efficiente nella gestione di grandi volumi di dati, poiché il confronto poteva coinvolgere un vasto elenco di città. L'ottimizzazione delle prestazioni è stata pertanto un'altra sfida significativa durante tutto il processo di sviluppo.

Durante l'implementazione, si sono riscontrate diverse difficoltà dovute alla mancanza di una documentazione chiara e completa per lo sviluppo e l'integrazione di nuovi componenti all'interno dell'ambiente Talend. Ciò ha portato a una svolta leggermente diversa rispetto all'obiettivo originale della tesi, optando per una soluzione che potesse risultare utile nella selezione del sistema per la gestione dei dati.

Questa nuova soluzione ha consentito una maggiore flessibilità nell'implementazione e ha ridotto le complessità tecniche precedentemente riscontrate. Sebbene il componente inizialmente progettato non sia stato effettivamente realizzato, si è deciso di adattare l'obiettivo presentando una soluzione alternativa che meglio soddisfa le esigenze e le risorse a disposizione per questo studio.

Questo adattamento e la capacità di riconsiderare l'obiettivo iniziale sono esempi dell'importanza della flessibilità nel processo di sviluppo software quando si devono affrontare ostacoli imprevisti.

5. DQ ANALYZER

5.1 Descrizione del progetto

All'interno del contesto di questa ricerca, è stata concepita e realizzata una webapp dedicata alla gestione della qualità dei dati, contribuendo in modo sostanziale all'incremento dell'affidabilità e all'efficacia delle operazioni di gestione dati. **DQ Analyzer** si focalizza sull'analisi della qualità dei dati forniti in input, mettendo a disposizione degli utenti una serie di controlli standard di fondamentale importanza. Tra questi, il controllo dei valori nulli, mirato a valutare la completezza delle informazioni, e la deduplicazione, che assicura la coerenza dei dati. Tuttavia, la vera unicità di questa soluzione emerge attraverso l'integrazione di funzionalità avanzate, quali la validazione degli indirizzi e-mail e la correzione dei campi contenenti nomi di città. DQ Analyzer non si limita semplicemente a individuare e segnalare potenziali problematiche nei dati, ma offre un feedback dettagliato insieme agli strumenti necessari per correggere gli errori rilevati. Tale approccio, focalizzato sull'automazione degli algoritmi di correzione, rappresenta un notevole avanzamento nel miglioramento della qualità dei dati e nell'incremento dell'efficienza delle operazioni di gestione dati, con impatti positivi in svariati settori applicativi. Nelle sezioni successive di questa tesi, verranno esaminati in dettaglio il processo di sviluppo sottostante e le sfide affrontate per portare a compimento questo progetto, evidenziando le sue potenzialità e le applicazioni pratiche che ne derivano.

5.2 Analisi dei requisiti funzionali

L'analisi dei requisiti del sistema è una fase fondamentale per definire chiaramente gli obiettivi e le funzionalità del sistema e ciò che l'utente si aspetta dal sistema stesso.

I requisiti funzionali corrispondono alle funzionalità o ai servizi che il sistema deve fornire. Essi descrivono anche il comportamento del sistema a fronte di particolari input e come esso dovrebbe reagire in determinate situazioni.

I requisiti funzionali individuati per il sistema DQ Analyzer sono i seguenti:

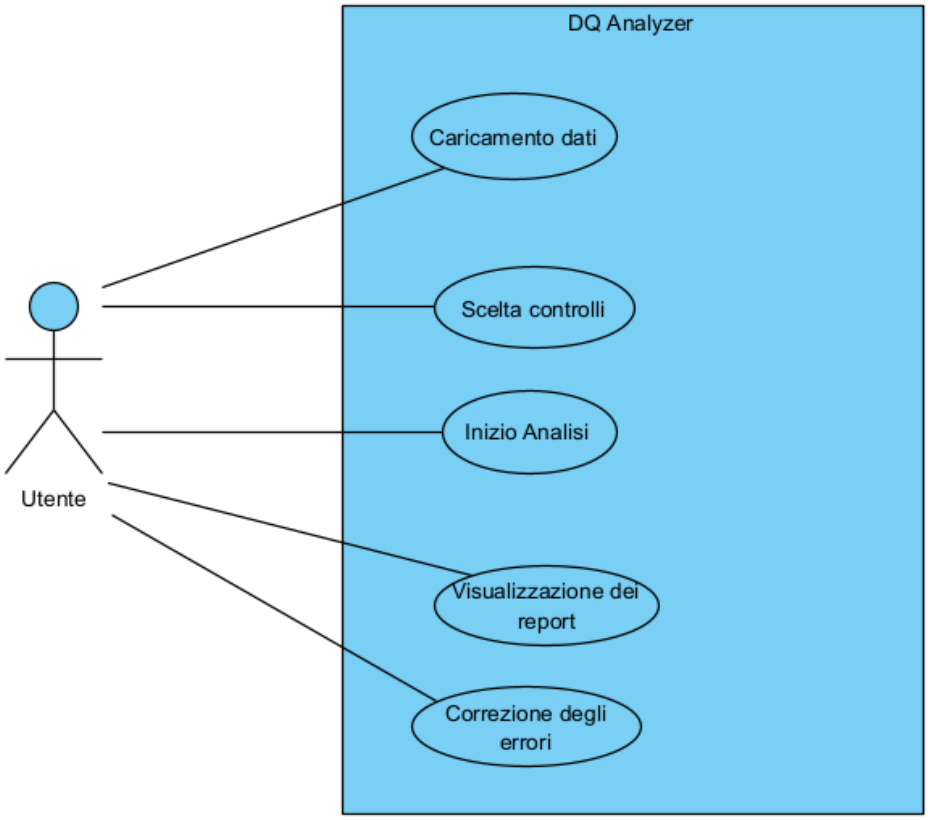
REQUISITO	DESCRIZIONE
RF01-Caricamento dei dati	Il sistema gestisce il caricamento dei dati da parte dell'utente, permettendo di effettuare l'upload di file o la connessione ad un database.
RF02-Controllo degli elementi nulli	Il sistema gestisce il controllo degli elementi nulli sui campi selezionati dall'utente.
RF03-Controllo degli elementi duplicati	Il sistema gestisce il controllo degli elementi duplicati sui campi selezionati dall'utente.
RF04-Validazione delle email	Il sistema gestisce la validazione delle email sui campi selezionati dall'utente.
RF05-Controllo della sintassi degli elementi	Il sistema gestisce il controllo della sintassi sui campi selezionati dall'utente, permette di effettuare il confronto degli elementi con elementi presenti in un altro file caricato dall'utente o messo a disposizione dal sistema.
RF06-Visualizzazione del report	Il sistema gestisce la visualizzazione del report dell'analisi effettuata sui dati dell'utente mostrando i risultati ottenuti in maniera chiara e leggibile.
RF07-Correzione degli errori rilevati	Il sistema permette all'utente di correggere gli errori rilevati.
RF08-Classificazione falsi positivi	Il sistema permette all'utente di classificare gli errori rilevati come falsi positivi, in tal modo questi non saranno considerati come errori nelle analisi

	seguenti.

5.3 Casi d'uso e scenari

I casi d'uso descrivono le interazioni tra gli attori e il sistema, gli attori rappresentano un ruolo all'interno del sistema.

I diagrammi dei casi d'uso descrivono gli attori in relazione ai casi d'uso attivati da essi.



Nome	**Caricamento Dati**
ID	CU1
Breve Descrizione	L'utente carica i suoi dati all'interno del sistema tramite l'upload di un dataset.

Attori Primari	Utente
Attori Secondari	Nessuno
Precondizioni	Esistenza deli file
Sequenza Principale degli eventi	1. Il caso d'uso inizia quando l'utente accede alla schermata per il caricamento 2. L'utente carica il file 3. Il sistema restituisce una schermata con l'esito del caricamento
Post-Condizioni	File caricato con successo
Sequenza Alternativa degli Eventi	File non trovato

Nome	Scelta dei controlli
ID	CU2
Breve Descrizione	L'utente vuole riconoscere un monumento scattando o caricando una foto
Attori Primari	Utente
Attori Secondari	Nessuno
Precondizioni	Nessuna
Sequenza Principale degli eventi	1. Il caso d'uso inizia quando l'utente accede alla schermata della scelta dei controlli 2. Il sistema fa visualizzare l'insieme dei campi e dei controlli possibili da effettuare. 3. L'utente sceglie i controlli da effettuare su ogni campo
Post-Condizioni	L'utente visualizza correttamente i

	controlli scelti
Sequenza Alternativa degli Eventi	Nessun campo trovato all'interno del dataset

Nome	Inizio analisi
ID	CU3
Breve Descrizione	L'utente inizia l'analisi di qualità sui suoi dati
Attori Primari	Utente
Attori Secondari	Nessuno
Precondizioni	L'utente ha scelto i controlli da effettuare
Sequenza Principale degli eventi	1. Il caso d'uso inizia quando l'utente sceglie di iniziare l'analisi 2. Il sistema effettua l'analisi dei dati e mostra il report dell'analisi
Post-Condizioni	Schermata di report visualizzata correttamente
Sequenza Alternativa degli Eventi	

Nome	Visualizzazione dei dettagli
ID	CU4
Breve Descrizione	L'utente visualizza i dettagli dell'analisi
Attori Primari	Utente

Attori Secondari	Nessuno
Precondizioni	Analisi avvenuta con successo
Sequenza Principale degli eventi	1. Il caso d'uso inizia quando l'utente vuole visualizzare i dettagli dell'analisi 2. Il sistema restituisce la schermata con i dettagli dell'analisi, mostrando i record analizzati e quelli errati
Post-Condizioni	Schermata visualizzata correttamente
Sequenza Alternativa degli Eventi	

Nome	Correzione degli errori
ID	CU5
Breve Descrizione	L'utente vuole correggere gli errori evidenziati dall'analisi
Attori Primari	Utente
Attori Secondari	Nessuno
Precondizioni	Nessuna
Sequenza Principale degli eventi	1. Il caso d'uso inizia quando l'utente vuole correggere gli errori risultanti dall'analisi. 2. Il sistema effettua una correzione automatica degli errori.
Post-Condizioni	Correzione avvenuta con successo
Sequenza Alternativa degli Eventi	Impossibile effettuare la correzione

5.4 Mockup

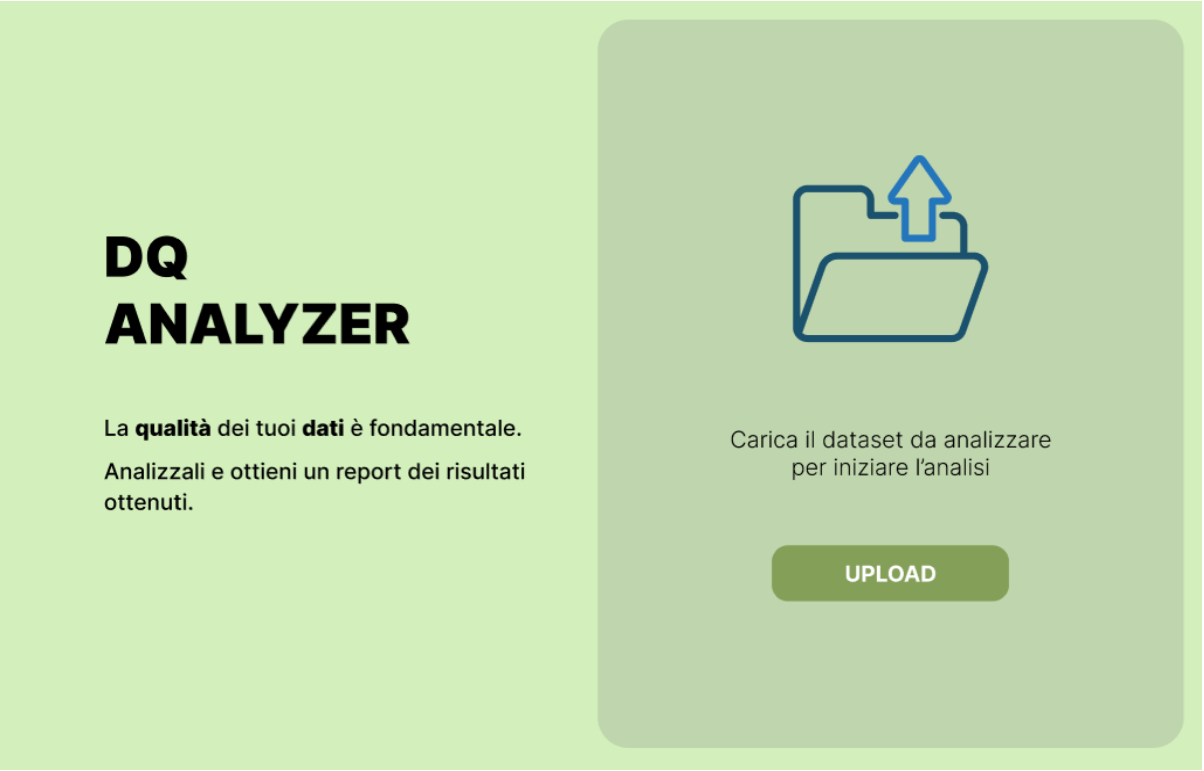
La fase successiva è stata lo sviluppo dei mockup dell'applicazione che hanno svolto un ruolo cruciale nella fase di progettazione e sviluppo. I mockup sono rappresentazioni visive dei vari aspetti e delle funzionalità chiave di un'applicazione, concepiti per illustrare in modo chiaro e tangibile l'aspetto e l'esperienza utente della soluzione.

I mockup hanno permesso di tradurre le idee e i requisiti funzionali in un formato visivo, in modo tale da poter valutare la disposizione degli elementi nell'interfaccia utente, la navigabilità tra le diverse schermate e la coerenza del design.

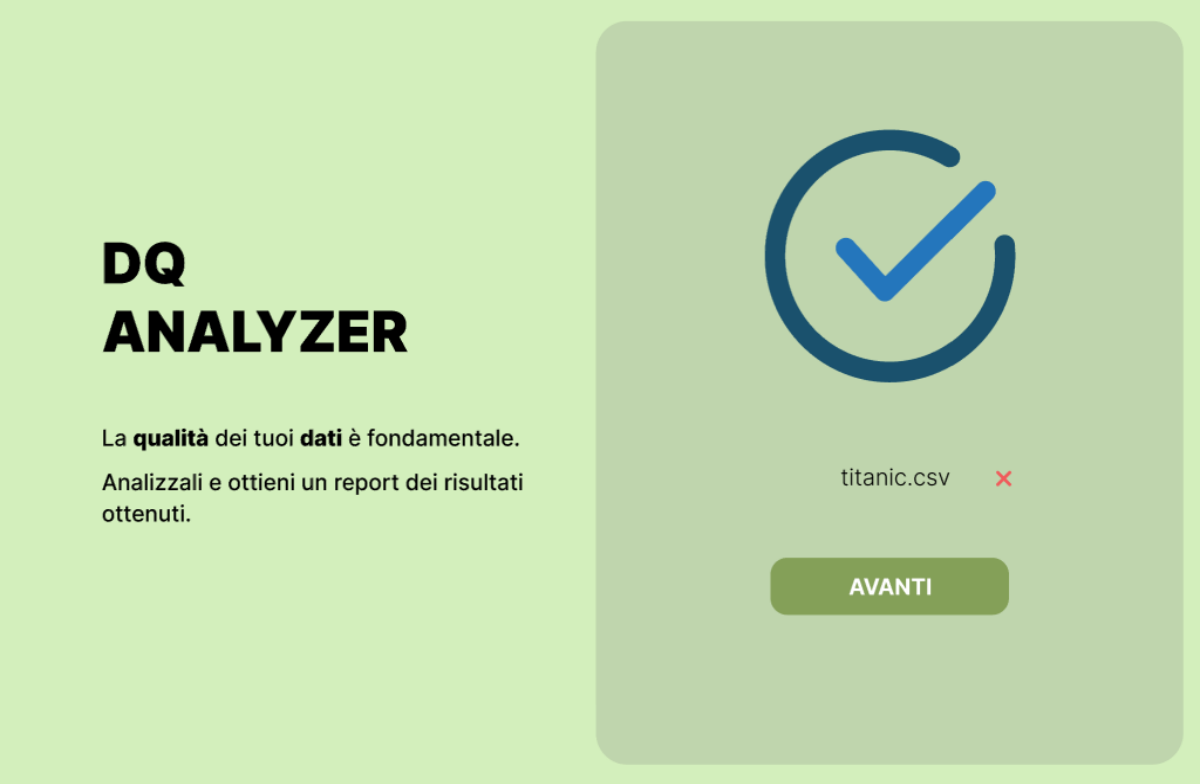
Di seguito verranno illustrate le diverse schermate create in fase di prototipazione:

- Schermata Home: presenta una breve descrizione dell'applicazione e permette all'utente di interagire iniziando l'analisi

- Schermata di upload: la prima cosa da fare per analizzare la qualità dei dati è caricare il proprio dataset, l'utente può quindi effettuare l'upload mediante questa schermata.

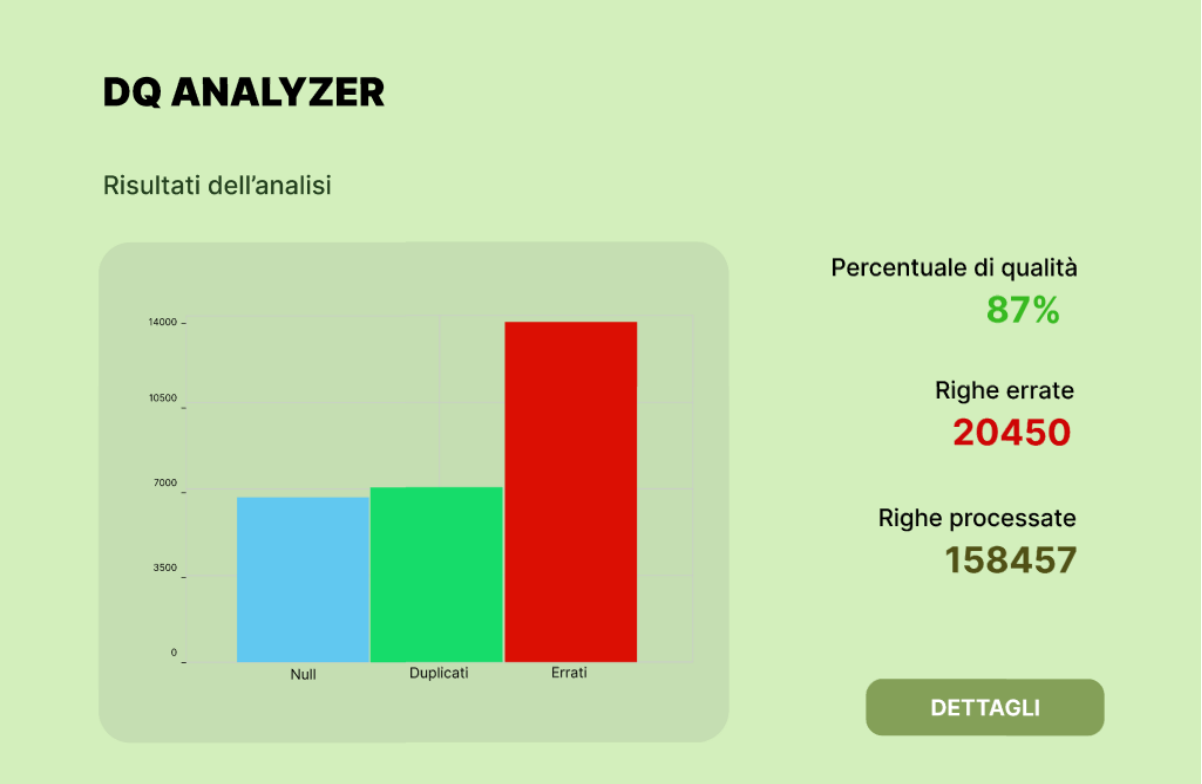


- Schermata di feedback: in seguito all'upload viene fornito un feedback visivo mostrando il successo del caricamento, questo è fondamentale in un'interrfaccia utente in quanto permette all'utente di capire in che stato si trova il sistema e quello che succede in seguito alle azioni compiute.

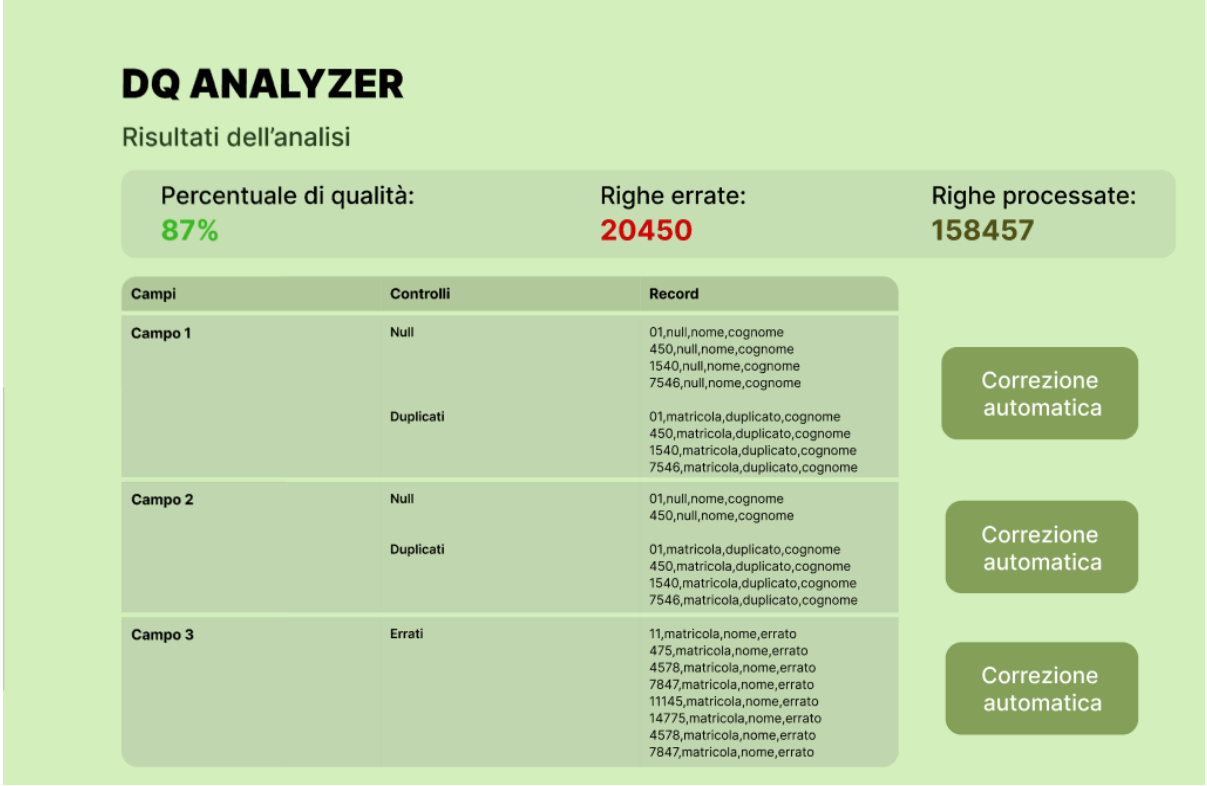


- Schermata per la scelta dei controlli: in questa schermata vengono mostrati i campi presenti nel dataset inserito dall'utente e su di essi i vari controlli che si possono effettuare con DQ Analyzer, è possibile quindi iniziare l'analisi di qualità.

- Schermata di monitoraggio e report: tramite questa schermata è possibile visionare i risultati dell'analisi con la percentuale di qualità raggiunta, l'utente può anche visualizzare i dettagli relativi all'analisi.



- Schermata dei dettagli:in questa schermata l'utente visualizza i dettagli dell'analisi effettuata, con i record che sono risultati errati e vi è la possibilità di effettuare la correzione automatica di tali errori.



5.5 Sviluppo

Per lo sviluppo è stato utilizzato il framework Yii 2, un framework open source che sfrutta il linguaggio PHP, pensato per sviluppare applicazioni web scalabili ed efficienti.

E' basato sull'architettura MVC (Model-View-Controller) che permette di separare la logica di presentazione dei dati dalla logica di business:

- Model (Modello): Il Model rappresenta la parte dell'applicazione responsabile della gestione dei dati e della logica di business. In Yii, i modelli sono rappresentati come classi PHP che mappano tabelle del database o oggetti di dati. Yii offre uno strumento chiamato ORM (Object-Relational Mapping) che semplifica l'interazione con il database. I modelli contengono metodi per accedere e manipolare i dati, consentendo agli sviluppatori di definire le regole di validazione dei dati e le relazioni tra i dati in modo dichiarativo.

- View (Vista): La View è responsabile della presentazione dei dati all'utente. In Yii, le viste sono file di template, spesso scritti in PHP, che contengono il codice HTML e incorporano dati dai modelli per creare la pagina visualizzata dall'utente. Yii supporta anche la creazione di widget riutilizzabili per semplificare la creazione di componenti di interfaccia utente complessi.

- Controller (Controllore): Il Controller gestisce le richieste degli utenti, coordina le interazioni tra il Modello e la Vista e decide quale azione deve essere eseguita in base all'URL richiesto. In Yii, i controller sono classi PHP che contengono azioni, ciascuna delle quali è una funzione responsabile di una specifica operazione. Yii utilizza il concetto di "routing" per mappare le URL alle azioni dei controller. Ad esempio, una richiesta all'URL "/site/start" potrebbe essere gestita dal controller "SiteController" e dall'azione "start".

Il flusso di lavoro tipico in un'applicazione Yii che segue il pattern MVC è il seguente:

1. L'utente fa una richiesta al server, ad esempio, visitando una pagina web.
2. Il sistema di routing di Yii determina quale controller e azione devono gestire la richiesta in base all'URL.
3. Il controller esegue l'azione corrispondente, interagendo con i modelli per recuperare o modificare dati secondo necessità.
4. Il controller seleziona la vista corretta, passando i dati necessari.
5. La vista elabora i dati e genera l'output HTML da restituire all'utente.
6. La pagina HTML generata viene inviata al browser dell'utente, che la visualizza.

Per lo sviluppo di DQ Analyzer, non risultava necessario avere dei model legati alle tabelle dei db, in quanto l'obiettivo dell'app è di far effettuare l'upload dei dati all'utente; pertanto, è stato creato un unico model per l'upload: CsvUploadForm.php che contiene i dettagli dell'upload e la funzione che gestisce l'upload.

Per quanto riguarda il controller, è stato utilizzato il SiteController, in cui sono state implementate diverse azioni per collegare le viste e funzioni per i controlli specifici da effettuare sui dati.

Le funzioni più rilevanti sono quelle che gestiscono i controlli sui dati: controlloNull verifica quanti e quali elementi all'interno del campo analizzato sono nulli e restituisce sia un report numerico che i record corrispondenti agli elementi nulli; controlloDuplicati verifica quanti e quali elementi all'interno del campo analizzato sono duplicati e restituisce sia un report numerico che i record corrispondenti agli elementi duplicati; validaEmail si occupa di verificare se gli elementi corrispondenti al campo analizzato corrispondono a email valide o meno, anche in questo caso viene fornito in output un report numerico e l'insieme dei record contenenti le email non valide; cleansingCittà esamina le città presenti all'interno del campo analizzato, le confronta con un file dell'archivio Istat che contiene tutte le città italiane corrette e fornisce in output un report numerico e l'insieme delle città che risultano errate correlate con una possibile correzione.

Le viste sviluppate riguardano la visualizzazione dei dati, quindi sono state sviluppate le viste per la visualizzazione dei risultati, quelle per i dettagli dei risultati che fornissero informazioni sui record errati e quelle per l'upload dei dati.

6. SPERIMENTAZIONE

Per la fase di sperimentazione è stato utilizzato un dataset contenente informazioni geografiche quali Indirizzo, CAP, Comune, Provincia e Regione. Il dataset è popolato da 15892 record.

Per la sperimentazione è stata valutata la qualità del dataset in questione utilizzando le misure di qualità dei dati previste dalle norme ISO/IEC e applicabili al dataset.

I risultati ottenuti mostrano che il dataset ha già un'alta qualità dei dati, in quanto non vi sono campi nulli o duplicati (Figura 7).

Per il campo Comune la qualità risulta migliorabile, infatti utilizzando le analisi effettuate dallo strumento sviluppato è stato possibile individuare che il 98.6% dei comuni soddisfa la qualità di accuracy, compliance e credibility; per la restante parte è possibile attuare dei miglioramenti della qualità.

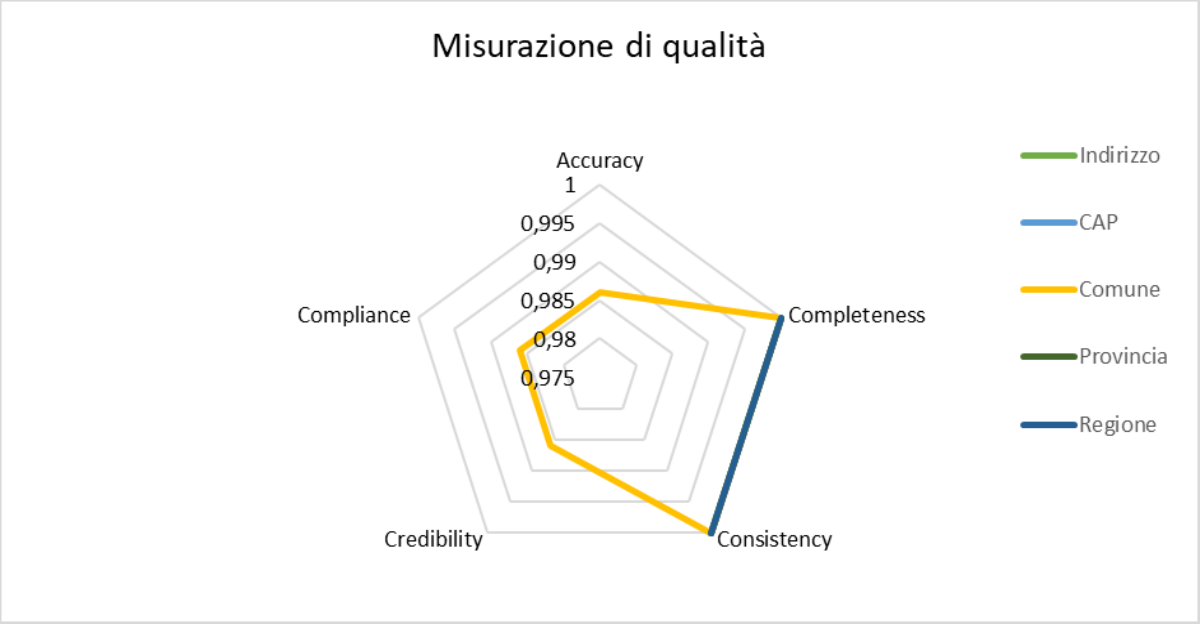


Figura 7: Misurazione della qualità

Quindi sono state effettuate delle correzioni al dataset, secondo i suggerimenti forniti dall'app e visualizzando i risultati sono stati trovati dei falsi positivi che sono stati classificati come tali in modo tale da permettere al sistema di non classificarli più come tali, per

le altre correzioni invece è stata effettuata la correzione secondo i suggerimenti forniti.

In seguito alle correzioni effettuate è stata nuovamente effettuata l'analisi di qualità (Figura 8) ed è stato rilevato un netto miglioramento, in quanto tutti i comuni sono risultati conformi alle caratteristiche di qualità, il che risulta essere un risultato soddisfacente per gli obiettivi preposti.

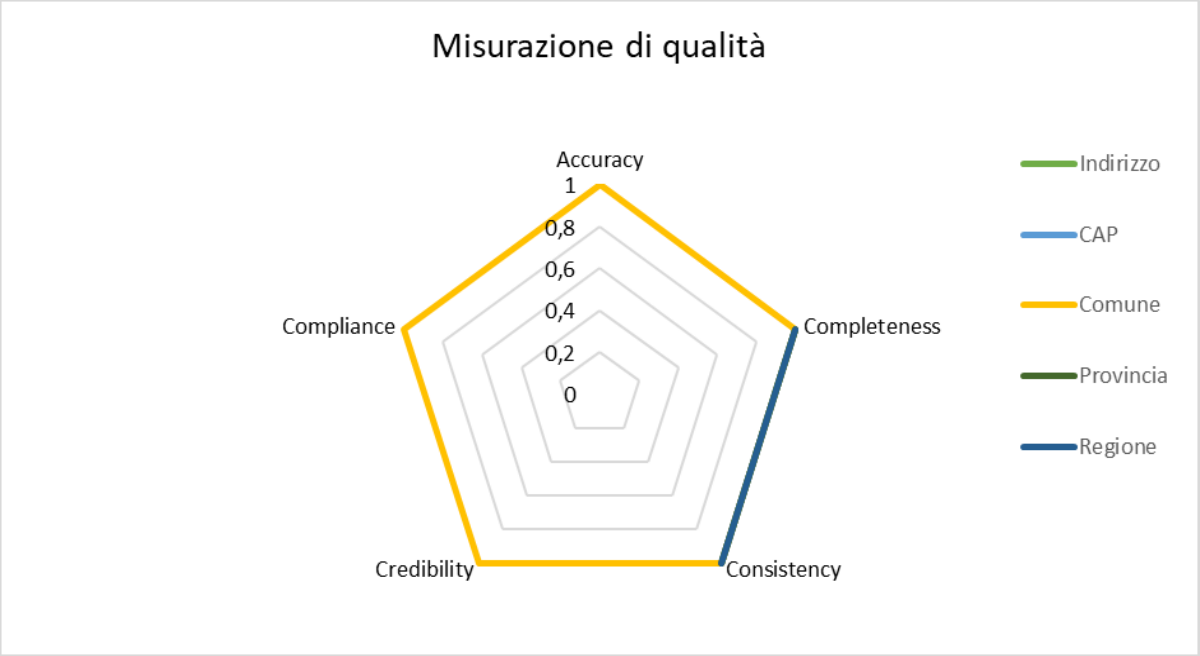


Figura 8: Misurazione di qualità in seguito alle correzioni

7. CONCLUSIONI

In conclusione, questa tesi ha affrontato l'importante tematica della gestione della qualità dei dati nei contesti aziendali, considerando il crescente volume di dati che le organizzazioni devono gestire. L'obiettivo iniziale era quello di analizzare il mercato dei sistemi software per la qualità dei dati, evidenziando punti di forza e debolezza delle soluzioni esistenti e proponendo un software adatto come punto di partenza per un sistema che soddisfacesse le esigenze degli utenti.

Tuttavia, il corso dello studio ha portato a un significativo cambiamento di prospettiva. La sfida di estendere la soluzione selezionata ha portato a una nuova direzione di ricerca, che si è concentrata sulla fornitura di un modello guida per la scelta degli strumenti più appropriati per la gestione della qualità dei dati in azienda. Inoltre, la tesi ha sviluppato un prototipo di componente software per la gestione del controllo di qualità dei dati, dimostrando la fattibilità dell'implementazione da zero.

Il principale obiettivo della tesi è quindi quello di contribuire all'ottimizzazione della gestione della qualità dei dati nelle organizzazioni. Ciò è stato realizzato fornendo un modello di guida per la selezione di strumenti adeguati e dimostrando la sua applicazione pratica attraverso lo sviluppo del prototipo. Questo approccio consentirà alle organizzazioni di effettuare scelte più efficaci e di migliorare la qualità dei loro dati, un elemento cruciale per il successo aziendale nell'era dell'informazione.

In un contesto in cui l'accuratezza e l'affidabilità dei dati sono essenziali per evitare errori, danni economici e reputazionali, l'analisi e il miglioramento continuo della qualità dei dati rappresentano una priorità.

La tesi ha quindi fornito un quadro concettuale solido, basato sulla comprensione del concetto di Data Quality, per fornire alle aziende e alle organizzazioni che basano le loro attività sulla gestione dei dati un modello guida solido per la scelta della soluzione da impiegare e ha mostrato la possibilità alternativa di poter implementare una soluzione adatta e personalizzata in base alle proprie esigenze impiegando risorse limitate.

8. LAVORI FUTURI

Gli obiettivi preposti all'inizio dello studio sono stati raggiunti con risultati soddisfacenti, ovviamente le valutazioni dei prodotti sul mercato necessitano di un continuo aggiornamento; pertanto, è necessario continuare a valutare le soluzioni presenti nello studio e gli eventuali aggiornamenti di pro e contro delle stesse.

Per quanto concerne il prototipo fornito, è stato un risultato soddisfacente in base alle risorse disponibili per lo studio, in futuro bisognerà attuare la logica implementata per il cleansing delle città ad altri campi rendendola meno specifica. Ad esempio, è possibile richiedere all'utente l'upload di un file di confronto per determinati campi o includere altri file nella soluzione.

Inoltre, bisognerà lavorare sulla connessione ai database sia relazionali che non relazionali, il lavoro effettuato si concentra sulla logica dei controlli, ma per renderlo una soluzione completa è necessario aggiungere la logica di connessione e attuare miglioramenti alla logica di reporting.

Infine, è possibile includere tecniche di machine learning quali il clustering e la classificazione per individuare eventuali outliers all'interno dei dati ed effettuare analisi sempre più automatiche.

9. BIBLIOGRAFIA

[1]: Wang, Richard Y., Mostapha Ziad, and Yang W. Lee. *Data quality*. Vol. 23. Springer Science & Business Media, 2006.

[2]: ISO/IEC 25000: 2014, Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE

[3]: J. Calabrese, S. Esponda, and P. Pesado, "Framework for Data Quality Evaluation Based on ISO/IEC 25012 and ISO/IEC 25024."

[4]: Altendeitering, Marcel and Tomczyk, Martin, "A Functional Taxonomy of Data Quality Tools: Insights from Science and Practice" (2022). Wirtschaftsinformatik 2022 Proceedings. 4.

https://aisel.aisnet.org/wi2022/business_analytics/business_analytics/4

[5]: ISO/IEC CD 25012, Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) – Data Quality Model

[6]: ISO/IEC CD 25024, Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of data quality

[7]: Talend: https://www.talend.com/it/

[8]: Gartner https://www.gartner.com/en

[9]: "Magic quadrant for data quality tools." *Gartner Inc* (2022).

10. RINGRAZIAMENTI

Se siete arrivati a questo punto quasi sicuramente ho finalmente una corona d'alloro in testa e stiamo festeggiando la conclusione di questo percorso.

È vero, solitamente i ringraziamenti sono più una formalità, la riuscita di un percorso universitario o di qualsiasi altro percorso infatti è merito di ognuno di noi; però è anche vero che in ogni parte della nostra vita ci saranno sempre persone che ci sosterranno e aiuteranno a raggiungere i nostri obiettivi nel migliore dei modi.

Per questo voglio ringraziarvi, ringrazio tutti voi presenti qui per aver contribuito (in un modo o nell'altro) al raggiungimento di questo obiettivo.

Parto col ringraziare tutta la mia famiglia per gli insegnamenti che mi avete dato e per il sostegno continuo in questi anni.

Ma in particolare grazie a te mamma, tu mi hai insegnato ad affrontare ogni ostacolo e che a volte è giusto chiedere aiuto a chi ci sta vicino. Grazie per la costanza nell'affrontare le cose che ci hai trasmesso.

Grazie a te papà, tu mi hai insegnato che a volte basta un po' di leggerezza e semplicità, che una risata può essere la soluzione per tutto e che non bisogna mai prendersi troppo sul serio.

Grazie a te Nico, questo traguardo è anche tuo, grazie per credere sempre in me e per avermi sempre spronata a fare di meglio.

Grazie a te Nonna Anna, per essere sempre presente, per insegnarci continuamente l'importanza del sostenersi e del volersi bene, a te che ci circondi di attenzioni e di amore dedico questo traguardo.

E adesso volevo ringraziare tutti gli amici che ci sono sempre stati, grazie ad ognuno di voi per avermi lasciato qualcosa del vostro carattere e per essere stati sempre pronti ad ascoltarmi e sostenermi.

Grazie al mio gruppo andriese, Mica, Fede, Fra Capo, Ste, Sabri, Lope, Ross per le chiacchierate, i pomeriggi al volver e le serate in piazza. Rimanete sempre una certezza.

Grazie a Franci, per trovare sempre del tempo per ascoltarmi e per condividere con me le tue esperienze, grazie per la bella amicizia e per essere sempre un motivo per cui vale la pena tornare.

Grazie a te Ale, che mi accompagni in ogni giornata, in ogni difficoltà, grazie perché mi hai insegnato il valore dell'amicizia e quanto questa sia fondamentale nei momenti migliori ma soprattutto in quelli peggiori. Grazie per essere così una persona vera e limpida.

Grazie a te Saretta, al nostro essere una famiglia, grazie perché anche se lontane non siamo mai poi così distanti, e soprattutto grazie per le mille esperienze vissute insieme e per trovare sempre un modo per farmi sentire la tua vicinanza.

E grazie a tutto il gruppo di Bari, siete diventati una vera e propria famiglia e sarà quello che più mi mancherà in questi due anni.

Grazie a te Flavia, per esserci stata dal primo momento, per non esserti mai allontanata, grazie per il sostegno che mi hai dato, per essere così simile a me e per capirmi anche solo con uno sguardo.

Grazie a Pet e al suo essere così intraprendente, mi hai insegnato che nella vita non bisogna mai accontentarsi di quello che abbiamo ma puntare sempre al meglio per noi.

Grazie a Gionno e agli esami, ai progetti insieme, alle chiacchierate, ma soprattutto al sushi e a tutti gli aperitivi al vergnano.

Grazie Dom per il tuo essere così Besos e per essere sempre pronto ad aiutare gli altri,

Grazie ad Angi, sei una persona splendida, grazie per tutti i tuoi consigli e gli aiuti in qualsiasi ambito, grazie soprattutto per il tuo essere sempre lì quando qualcosa non andava.

Grazie Moni per tutte le sclerate e i caffè sfogo pre-laurea.

Grazie alla house più visitata di Bari, in particolare grazie Flavius, Kuco e Alessio, per le infinite serate a casa vostra e per aver reso possibile la consegna di questa tesi.

Grazie a Vinci, il coinquilino più folle di sempre, con te non ci si annoia mai, grazie per averci sempre strappato un sorriso, mi mancherai.

E infine grazie a voi, i miei coinquilini attuali.

Grazie Mat per essere una persona così buona e gentile, anche se non lo sai neanche tu, grazie per avermi ascoltata e per avermi permesso di ascoltarti, grazie per esserci sempre.

Grazie Cri per tutte le volte che mi hai svegliata perché rimandavo le sveglie, per tutti gli esami preparati insieme, per l'amicizia che abbiamo e perché sei una persona fantastica.

Ci sarebbero mille altre cose da dire e tante altre persone da ringraziare; quindi, ringrazio semplicemente tutti voi per essere qui e nella mia vita. Vi voglio bene.