UNIVERSITA' DEGLI STUDI DI TORINO

MASTER di I LIVELLO in

ANALISI DATI PER LA BUSINESS INTELLIGENCE E DATA SCIENCE

ANNO ACCADEMICO 2024/2025

Titolo della tesi: **Gestione dei dati di progetto tramite sistema ETL: un caso applicativo nella progettazione ambientale**

Autore: Diego Cavallo

Abstract

Il presente lavoro di tesi documenta la progettazione e lo sviluppo di un sistema ETL (Extract, Transform, Load) per la gestione strutturata dei dati di progetto presso una società di consulenza operante nel settore dei servizi ambientali e dell'igiene urbana. Il contesto di partenza era caratterizzato da un patrimonio informativo significativo ma frammentato: circa novanta progetti tecnico-economici gestiti nel corso degli anni, ciascuno documentato in file Excel con strutture simili ma non uniformi, archiviati sul server aziendale senza alcuna forma di integrazione sistematica. Questa frammentazione rendeva impossibile effettuare analisi comparative tra progetti, estrarre indicatori aggregati o sfruttare i dati storici per supportare la stima dei costi nelle nuove commesse.

Il progetto nasce da un mandato aziendale specifico: valorizzare questo patrimonio informativo costruendo le fondamenta tecnologiche necessarie per abilitare, in prospettiva futura, analisi predittive e modelli di stima automatizzata. La soluzione sviluppata si articola in tre componenti principali: script Python per l'estrazione e la trasformazione dei dati dai file Excel, un database SQL Server come repository centralizzato, e un'interfaccia Microsoft Access che consente agli utenti non tecnici di operare il sistema senza interagire direttamente con il codice.

La fase di analisi ha rivelato che la principale sfida tecnica non risiedeva nell'importazione dei dati in senso stretto, bensì nella gestione della variabilità strutturale dei file Excel: fogli con nomi diversi,

colonne in posizioni variabili, celle di riferimento non standardizzate, e casi particolari legati a progetti multi-comune. Per affrontare questa eterogeneità è stato sviluppato un approccio basato su un file di configurazione centralizzato che mappa i metadati di ciascun progetto, consentendo agli script di adattare dinamicamente la logica di estrazione alle specificità di ogni file.

Lo sviluppo ha seguito un percorso iterativo, con una fase esplorativa iniziale che ha portato ad abbandonare soluzioni native di SQL Server, come il wizard di importazione e la funzionalità OPENROWSET, a causa di problemi di instabilità con file di grandi dimensioni. La scelta di Python come tecnologia di elaborazione si è rivelata determinante per la robustezza e la flessibilità del sistema. Gli script implementano logiche differenziate per ciascuna tipologia di dato: allineamento dinamico delle colonne per i servizi, taglio basato su riferimento di cella per manodopera e mezzi, ricerca testuale per gli indici economici del quadro generale. La pulizia e la validazione dei dati importati sono delegate a stored procedure SQL Server, mantenendo gli script Python focalizzati sulla sola fase di estrazione e caricamento.

Il sistema è attualmente operativo e in uso nella gestione ordinaria dei progetti aziendali. L'elaborazione massiva dell'intero archivio storico richiede circa quindici minuti, mentre l'importazione incrementale di un singolo nuovo progetto avviene in pochi secondi attraverso l'interfaccia Access. Il lavoro identifica con trasparenza le criticità residue, in particolare la gestione delle credenziali in ambienti multi-utente e l'assenza di funzionalità di export, e delinea un percorso evolutivo che prevede, a medio termine, l'integrazione con strumenti di business intelligence e, a lungo termine, lo sviluppo di modelli predittivi per la stima automatizzata dei costi di appalto.

Dal punto di vista metodologico, il progetto illustra come un approccio pragmatico e iterativo, che privilegia la semplicità e l'affidabilità rispetto all'eleganza teorica, possa produrre risultati concreti in un contesto aziendale reale, dove i dati da gestire sono imperfetti per natura e le soluzioni devono essere adottabili anche da utenti non specializzati.