Master universitario in: “ANALISI DATI PER LA BUSINESS INTELLIGENCE”

A.A. 2015/2016

Titolo Tesi: Il progetto di valorizzazione del patrimonio informativo della banca dati sulla finanza territoriale Ires Piemonte: progettazione e implementazione di un processo di Data Quality e Data Integration applicato ai dati di bilancio comunali di fonte Ministero dell’Interno.

Autore: Francesca Nazzi

Abstract: Il progetto nasce con lo scopo di ampliare la fruibilità dei dati raccolti e dei risultati di ricerca dell’Osservatorio per la Finanza Territoriale dell’IRES, al fine di valorizzare l’aspetto pubblico e la qualità della ricerca svolta negli uffici dell’ente, e rendere significativi, ossia intelligibili, dati di bilancio (consuntivo e preventivo), analisi fiscali e previsioni finanziarie dei Comuni e delle amministrazioni sovracomunali del Piemonte.
Si tratta di un lavoro di progettazione e pulizia dell’archivio dati dell’IRES, di fonte Ministero dell’Interno, ed è il primo step di un progetto di sviluppo di una piattaforma web, che prevede la creazione di un ampio e articolato Data Base (DB) in back-end che contenga non solo i dati non elaborati, coerentemente strutturati, ma un corredo di funzioni di calcolo che rispondano alle richieste di interpretazione e analisi più frequenti dell’utenza dell’Ires.
Nel concreto è stata svolta una parte preliminare dell’intero processo di Data Quality e Data Integration, che richiederebbe una disponibilità di tempo di gran lunga maggiore. Il mio lavoro si è articolato in due fasi che sono venute definendosi man mano che si chiarivano i dettagli e alcune lacune dell’archivio dati.
In primo luogo mi sono occupata del controllo della coerenza interna dei dati anagrafici dei comuni nazionali: conteggio dei missing e della completezza dei record, controllo della coerenza dei dati con le fonti esterne ufficiali, quali l’Istat.
In un secondo luogo è stato progettato un sistema di etichettatura e ridefinizione della modalità di caricamento dei nuovi dati in archivio.
Il software di lavoro è stato SAS Base, SAS Enterprise Guide.
Le difficoltà principali, che hanno determinato l’incompletezza del progetto soprattutto nella seconda parte, concernono la quantità e l’eterogeneità dei file. Infatti i dati provengono dal “Certificato di rendiconto al bilancio” che ogni anno gli enti comunali e provinciali consegnano al Ministero dell’Interno. Questi documenti sono però soggetti alle continue variazioni legislative e ciò comporta che di anno in anno il numero di voci, quindi di variabili, aumenti o diminuisca anche in modo sostanziale. Ciò rende laborioso e lento scrivere un programma che processi in automatico tutti i file tenendo conto di tutte le eccezioni.
Al momento di consegna del lavoro restano dei passi da compiere:
trovare una funzione efficace per rinominare più di 500 variabili sul dataset di partenza;
scrivere un programma di ridefinizione del formato di più di un centinaio di variabili per ogni file;
confrontare i campi delle voci di bilancio che hanno mantenuto lo stesso codice univoco pur subendo variazioni normative significative, e segnalare la variazione del sistema di etichettatura.