

Relazione Project Group

Bilancio di Genere

**Studenti**

*Andrea Perlini*

*Donoval Candolfi*

*Alessandro Tullio*

**Obiettivo**

L’obiettivo del progetto è quello di poter visionare l’andamento della parità di genere all’interno della nostra Università generando una dashboard interattiva in cui visionare i dati, in quanto deve essere garantita la rappresentanza e promossa la parità di genere all’interno degli organi universitari.

Per poter effettuare questo progetto , ci siamo avvalsi di uno strumento chiamato Pentaho, cioè un ETL ,cui acronimo significa Estrazione Trasformazione e Caricamento dati , che ci ha permesso di processare diverse fonti di dati , poterle integrare fra loro e analizzarle in un secondo momento.

**Strumenti Utilizzati**

* Pentaho Data Integration
* Pentaho BI Server
* Database Microsoft SQL Server
* Python

**Pentaho Data Integration**, detto anche Kettle, è uno strumento fornito da Pentaho, per poter migrare dati, fornire integrazioni tra applicazioni basate sui dati e creazione di banche dati orientate all’analisi.

È formato dai seguenti componenti : Spoon per effettuare un disegno grafico dei passi dell’ETL, Pan per l’esecuzione delle Trasformazioni da linea di comando, Kitchen per l’esecuzione dei Job da linea di comando, e Carte per l’esecuzione remota dei processi.

**Pentaho BI Server** è la componente di Pentaho che si occupa della presentazione dei dati, accedendo tramite browser si possono realizzare report o grafici, o rendere disponibili sul web delle applicazioni realizzate con gli altri tool di Pentaho .

L’utilizzo è molto semplice, si tratta di un’ applicazione web che con un menu ed alcuni wizard consente la creazione, la gestione e l’utilizzo dei contenuti.

Al suo interno ci sono molteplici componenti nominati CTools, come per esempio :

* **CDF**: Community Dashboard Framework
* **CDA**: Community Data Access Layer
* **CDE**: Community Dashboard Editor
* **CCC**: Community Charting Components
* **CGG**: Community Graphics Generator

Il Database **Microsoft SQL Server**, come nostra Data Warehouse, ci ha permesso di leggere e analizzare i dati, elaborati e trasformati precedentemente.

Infine si è usato il linguaggio di programmazione **Python**, per poterci agevolare il lavoro di recupero dati, vista la limitata disponibilità di strumenti offerti da Kettle. Grazie a questo, siamo riusciti ad elaborare dati tabellari contenuti in Tag HTML o convertire formati diversi, esportandoli in un formato a noi desiderato.

**Data Source**

* Cineca - Cerca università
* Unimol – Ricerca Abilitati ASN

Da questi Data Source sono stati esportati dei dati tabellari.

Da Cineca abbiamo scaricato il contenuto in HTM ed è stato convertito poi in XLSX.

Riguardo i dati provenienti da Ricerca Abilitati ASN di Unimol ,si sono navigate tutte le tabelle sfruttando i tag Table facenti parte dell’HTML, esportando il tutto in XLSX.

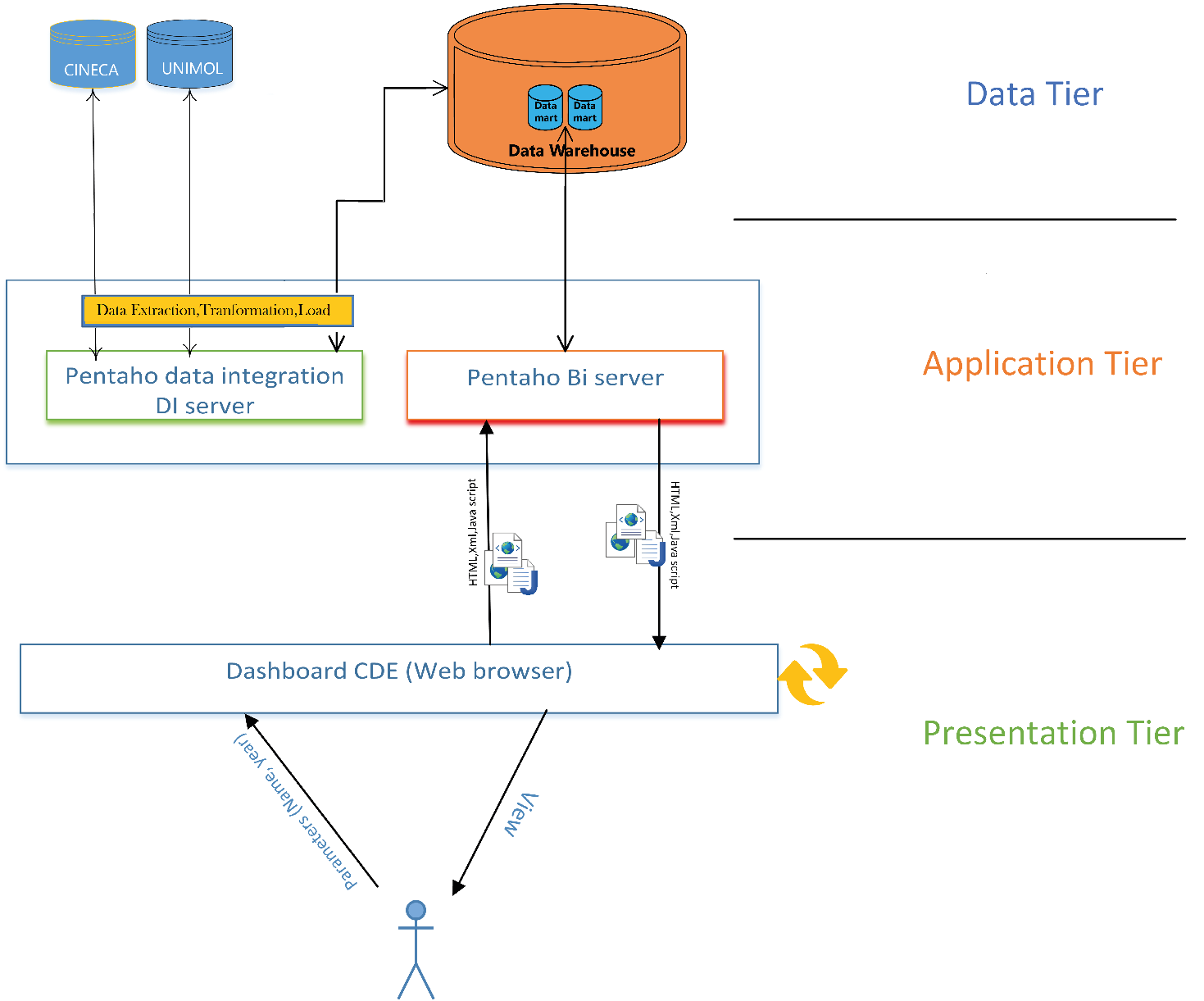
Ci è stato di supporto l’uso di Script in Python, che ci ha consentito di elaborare queste informazioni, dato che Kettle non aveva la possibilità di prendere in input file in formato HTM.

**Dataset Finale**

Il nostro Dataset è memorizzato in una tabella nel database Pentaho del nostro SQL Server.

Questa tabella contiene le tabelle prese in esame in questo progetto :

* Ateneo
* Facolta
* Ruolo
* SSD
* Afferenza
* Servizio\_prestato\_in\_altro\_ateneo
* UUID
* Cognome\_e\_Nome
* Genere
* Tornata
* Settore
* Fascia
* Q
* Inizio
* Fine



**Attività svolte**

In primo luogo ci siamo occupati del Data Source di Cineca, cioè il sito del Ministero dell’Istruzione, dell’Università e della Ricerca, che raccoglie le informazioni sul personale docente di tutte le università italiane.

Abbiamo preso in esame l’Università di Camerino, e ci siamo concentrati sulla funzione presente sul sito che permette di esportare la tabella in formato Excel.

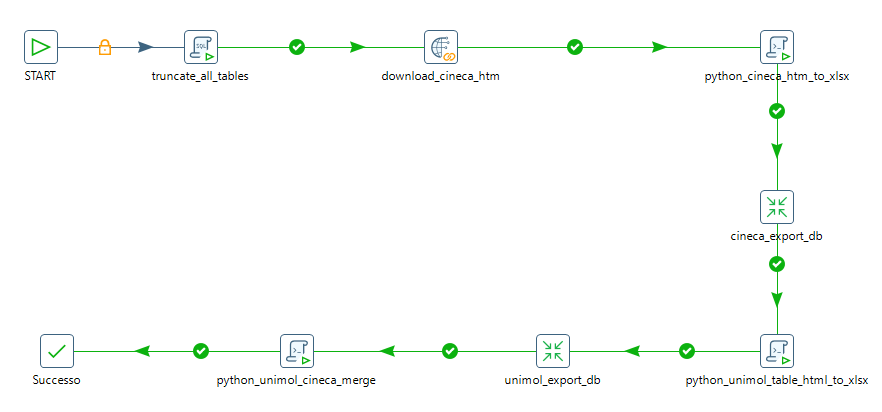
Per far ciò, si è proceduti a ispezionare le richieste che venivano scatenate dal browser web, ed è stato possibile risalire all’URL di download del file.

In secondo luogo, ci siamo concentrati sul Data Source di Unimol riguardante l’Abilitazione Scientifica Nazionale, che consente al personale docente di poter partecipare ai concorsi e poter essere assunti in un ambiente di istruzione.

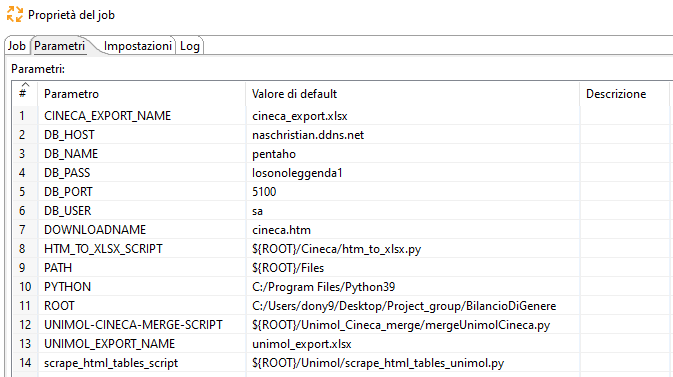
Abbiamo notato che la funzione di esportazione della tabella non consentiva di scaricare la tabella completa, ma soltanto quella presente nelle singole pagine.

Abbiamo iniziato il lavoro su Pentaho Data Integration per poter effettuare Job e Trasformazioni.

In figura, il job Main completo di tutte le trasformazioni.



Per agevolarci il tutto, si sono create delle variabili globali all’interno del job Main.

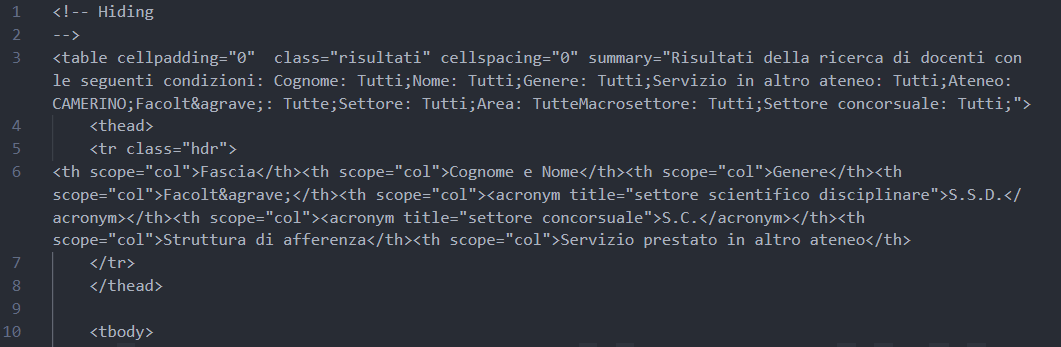


**Cineca**

Ci siamo occupati in primo luogo del Data Source di Cineca.

Nell’ambiente grafico Spoon, è stato aggiunto un oggetto di base denominato HTTP, presente in figura con il nome di “download\_cineca\_htm”, in cui abbiamo immesso l’URL di download del file.

Studiando la struttura del file, abbiamo riscontrato che non si presentava come file tabellare XLS, ma aveva la struttura di un file HTM.

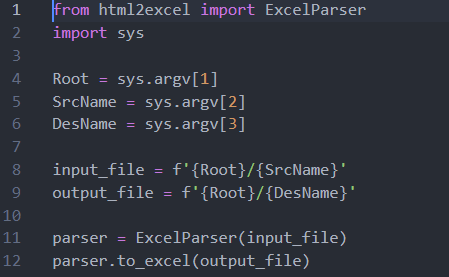


*File cineca.htm*

Da qui è nata l’idea di scaricarlo con l’estensione corretta in HTM.

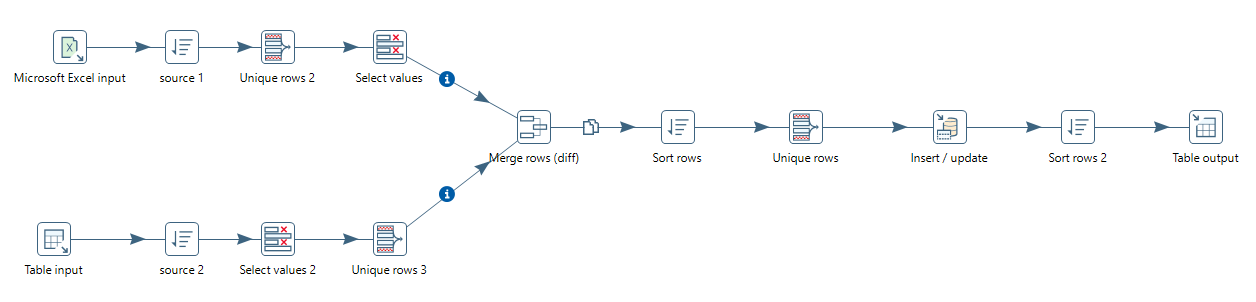
Un limite di Pentaho Data Integration è quello di non poter prendere in input file in HTM, quindi si è studiata una soluzione alternativa.

Si è scelto di prendere in esame uno script in Python creato con il nome di “htm\_to\_xlsx” integrato nell’oggetto “python\_cineca\_htm\_to\_xlsx”, che ci ha permesso di effettuare la conversione in realtime di codesto file in un formato tabellare XLSX successivamente dato in Input a Pentaho.

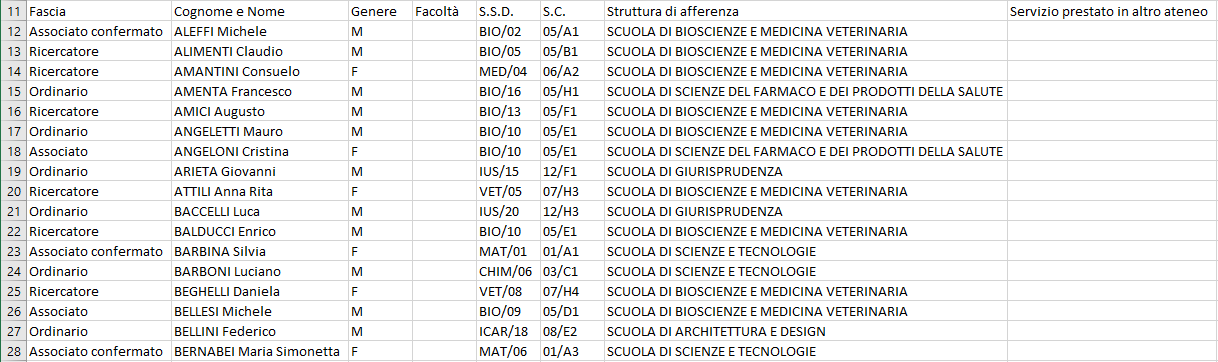


*Script htm\_to\_xlsx.py*

Poi si è predisposta una Trasformazione denominata “cineca\_export\_db” in grado di effettuare un Output di dati sul nostro database SQL Server, in una tabella precedentemente creata con le stesse colonne presenti nel file XLSX dal nome di “cinecaDocenti”.



*Trasformazione cineca\_export\_db.ktr*

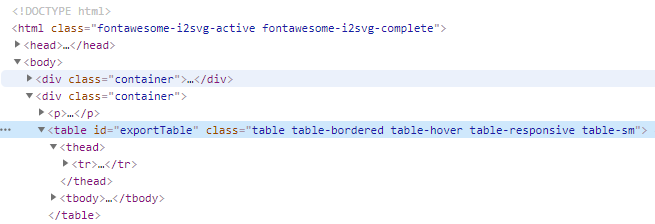


*File cineca\_export.xlsx*

**Unimol**

In secondo luogo siamo intervenuti sul Data Source di Unimol.

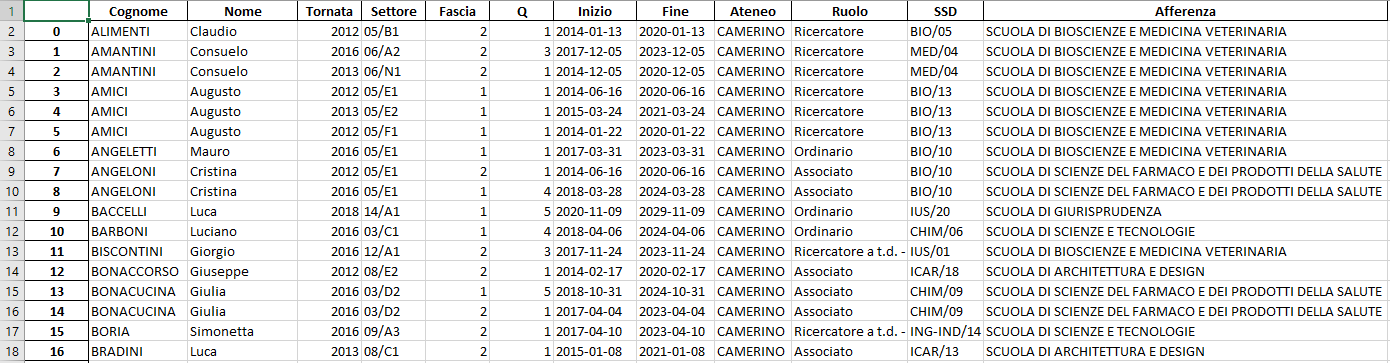
Per il problema sopra descritto, si è pensato di creare un ulteriore script in Python chiamato “scrape\_html\_tables\_unimol” e rappresentato in figura con l’oggetto “python\_unimol\_table\_html\_to\_xlsx”, che grazie alla libreria di Data Analysis “pandas”, ci ha permesso di scansionare in ogni singola pagina il tag HTML <table>.





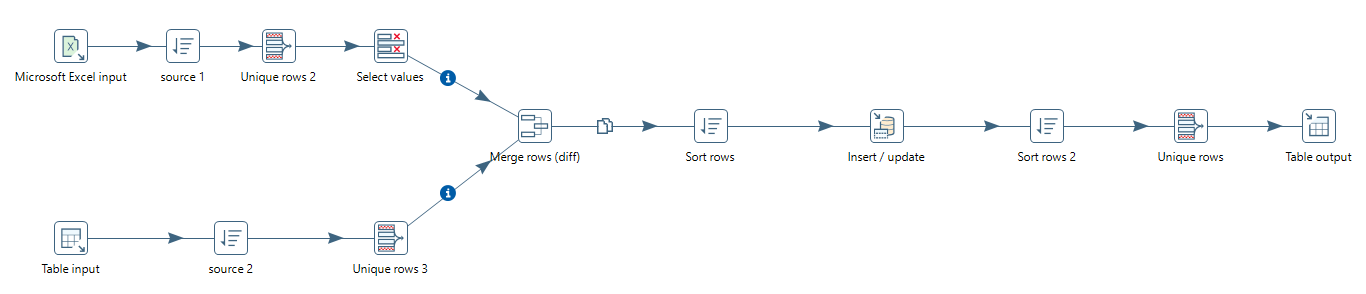
*Script scrape\_html\_tables\_unimol.py*

Di conseguenza è stato esportato in formato XLSX per poter essere analizzato da Pentaho.



*File unimol\_export.xlsx*

Si è predisposta una Trasformazione “unimol\_export\_db” , che esegue l’Output dei dati sul nostro database SQL Server, in una tabella precedentemente creata, con le stesse colonne presenti nel file XLSX dal nome di unimolDocenti.

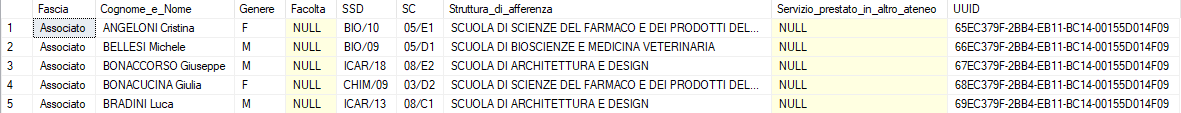


*Trasformazione unimol\_export\_to\_db.ktr*

A questo punto, si è iniziato a confrontare le singole tabelle.

Abbiamo notato un incompatibilità delle colonne, e l’assenza di informazioni chiave riguardante tutti i campi.

In dettaglio, non erano presenti in entrambe le tabelle le informazioni in grado di soddisfare l’obiettivo del progetto.



*Tabella cinecaDocenti*



*Tabella unimolDocenti*

**Unione delle tabelle**

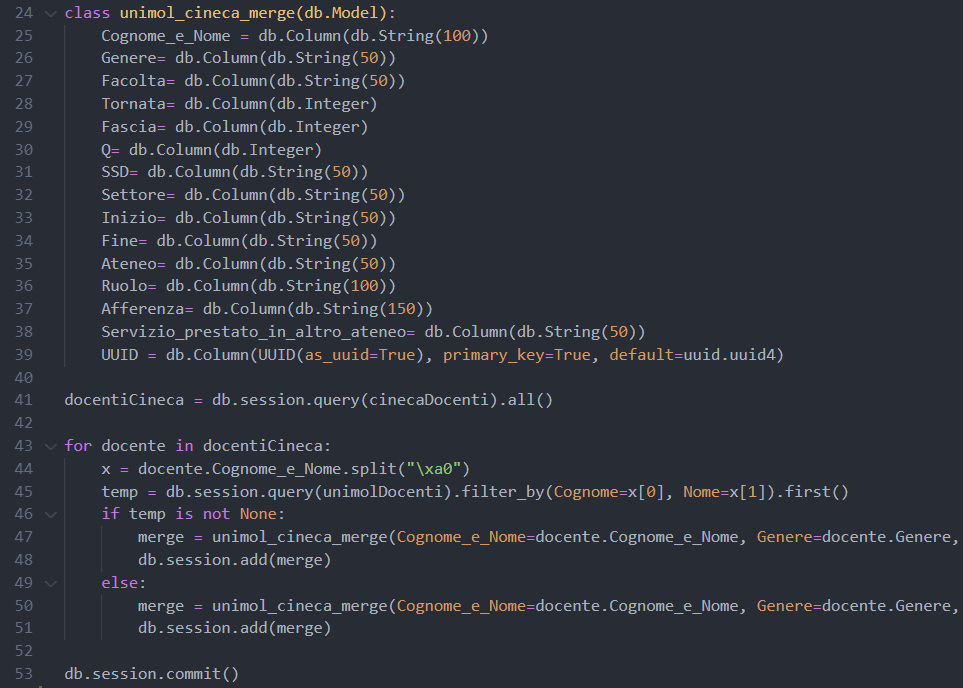
Si è pensata ad una soluzione che ci permettesse di poter unire le due tabelle presenti nel DB, mappando le informazioni mancanti ed eventualmente aggiungendole.

Abbiamo notato dei limiti di Pentaho Data Integration, perché non avevamo bisogno di una semplice Join in grado di aggiungere una colonna, ma avevamo la necessità di mappare ogni singola riga con i dati mancanti confrontando le tabelle.

Per far ciò, siamo ricorsi ad un ulteriore script in Python chiamato “mergeUnimolCineca” e rappresentato in figura con l’oggetto “python\_unimol\_cineca\_merge”, che ha semplificato e reso agevole il lavoro da svolgere.



*Script mergeUnimolCineca.py*



*Script mergeUnimolCineca.py*

Di seguito la tabella completa di tutte le informazioni inserite nel DB.



*Tabella unimol\_cineca\_merge*