Progetto Business Intelligence

Note da considerare:
1. Indicare che il pagamento è relativo ad un solo noleggio auto, come noleggio auto è relativo ad un unico pagamento

1. Descrizione dello scenario

Possiamo ipotizzare che il sistema del committente al momento sia molto arretrato, con l'utilizzo esclusivo di un DB relazionale per memorizzare le informazioni relative ad i noleggi delle automobili. Un business user per reperire le informazioni sui noleggi deve quindi formulare query sql molto complesse che difficilmente gli permettono di rintracciare con facilità le informazioni richieste.
La difficoltà nel formulare le interrogazioni rallenta i processi decisionali aziendali, che diventano macchinosi e poco produttivi.

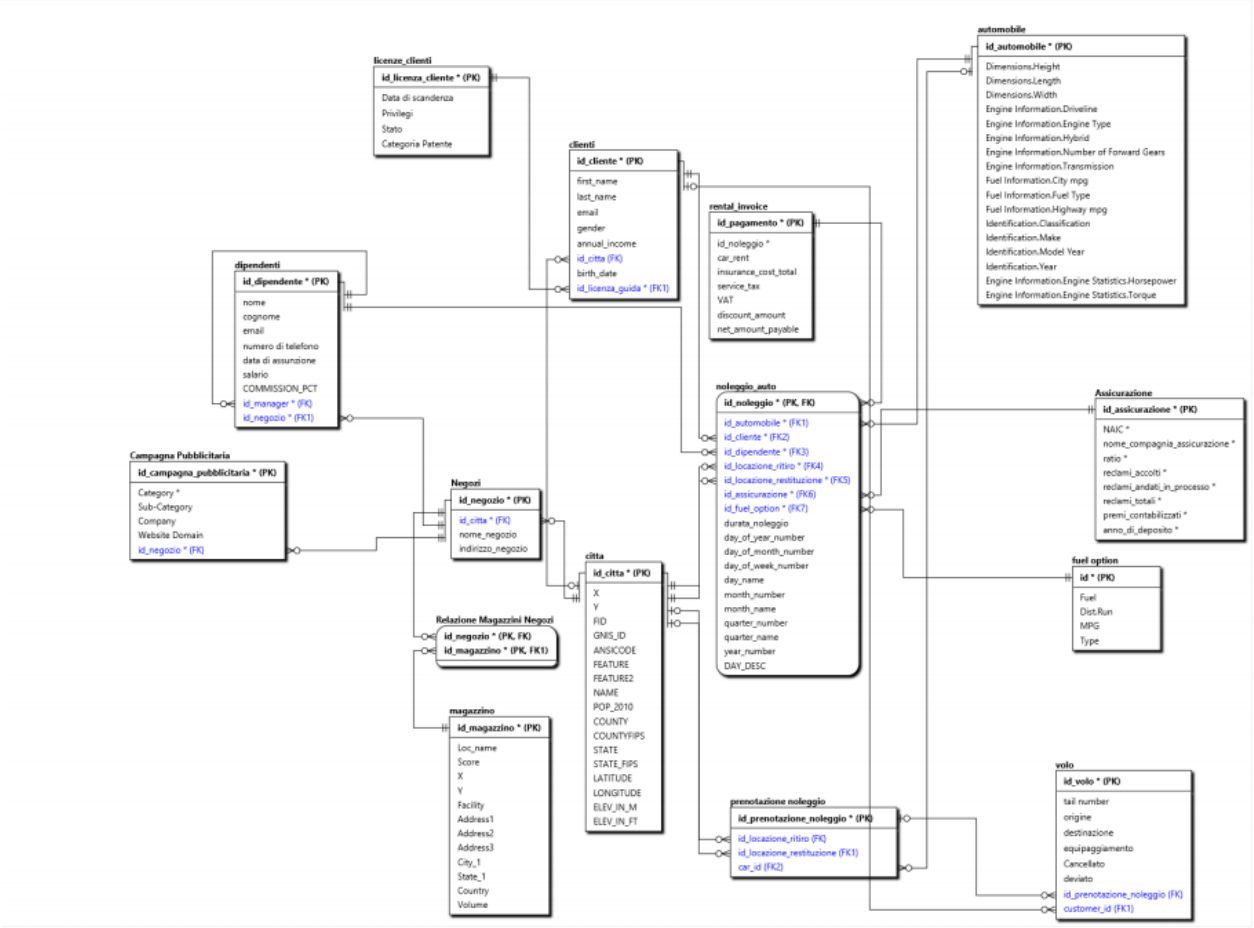
Le sorgenti dati disponibili sono le seguenti:

Cities_and_Towns_NTAD.csv	Contiene le informazioni su città, stato e contea utilizzate nel sistema. Quindi i luoghi in cui sono state ritirate o consegnate delle automobili.
automobile-insurance-company-complaint-rankings-beginning-2009-1.csv	Contiene le informazioni sulle compagnie assicurative che hanno sottoscritto i contratti di noleggio.
Merge - csv.com__60241c6adec19.csv	Sono memorizzate le informazioni sui clienti che hanno noleggiato un'automobile
driver-license-permit-and-non-driver-identification-cards-issued-as-of-august-30-2017-1.csv	Sono presenti le informazioni sulle licenze di guida dei clienti che hanno effettuato un noleggio.
martech2017.xlsx	Contiene le informazioni sulle campagne pubblicitarie effettuate dagli autonoleggi
MOCK_DATA (5).csv	Informazioni sui dipendenti dei vari autonoleggi
store.csv	Contiene le informazioni sugli autonoleggi in cui sono state noleggiato delle automobili

È presente, infine nel DB una tabella che esprime le relazioni di noleggio di un'automobile da parte di un cliente in un particolare autonoleggio.

2. Analisi e Riconciliazione delle fonti dati

Dopo una prima analisi delle fonti dati, abbiamo inserito tutto in un unico DataBase in modo da avere tutte le informazioni raccolte in un'unica locazione. A seguire abbiamo iniziato con la fase di analisi.



2.1 Ricognizione

Il dominio applicativo consiste in informazioni relative al noleggio di automobili ed eventualmente alla prenotazione di un noleggio.

Sono presenti informazioni riguardanti il cliente che ha sottoscritto il noleggio con riferimento alla sua città di provenienza ed alla sua licenza di guida.

Per ogni noleggio è indicata l’automobile che l’utente ha noleggiato, con indicata la marca, la categoria ed altre informazioni.

Sono riportati anche luogo di consegna e di ritiro dell’auto noleggiata.

Oltre a quanto detto, sono presenti informazioni sul dipendente che ha firmato il contratto di

noleggio ed il negozio per cui lavora con l'aggiunta delle informazioni anagrafiche e del manager a cui risponde.

Inoltre, sempre in relazione al noleggio abbiamo la data di firma del contratto di noleggio.

La tabella campagna pubblicitaria contiene le informazioni sulle campagne effettuate dalla concessionaria per pubblicizzare i propri servizi.

Per ogni negozio sono anche riportati i magazzini a lui collegati ed in cui sono smistate le varie automobili prima di essere portate nella concessionaria.

La tabella fuel option contiene informazioni relative allo stato del veicolo al momento della consegna, con informazioni sulla quantità di benzina nel serbatoio e sui chilometri percorsi.

Infine, è presente una tabella prenotazione che contiene le informazioni sulla prenotazione di un noleggio effettuata da un cliente che ha usufruito di un volo, e che quindi ha prenotato l'automobile nello stesso momento in cui ha prenotato il volo.

2.2 Normalizzazione

Negli schemi locali sono presenti alcune dipendenze funzionali non esplicitate direttamente, che sono:

- città → stato →Country, in Città

  Sono già presenti gli identificatori degli stati e dei country, e corrispondono ad i valori STATE_FIPS, COUNTRY_FIPS. Sono degli identificatori nazionali di stati e country

- città → stato →Country, in magazzino.

  Avendo normalizzato città non è necessario normalizzare anche magazzino in quanto basta sostituire al nome della città con il suo ID. Abbiamo fatto questa sostituzione inserendo un passo di join su nome e poi lasciando solo l'id.

  È stata aggiunta in questo modo una chiave esterna tra magazzino e città

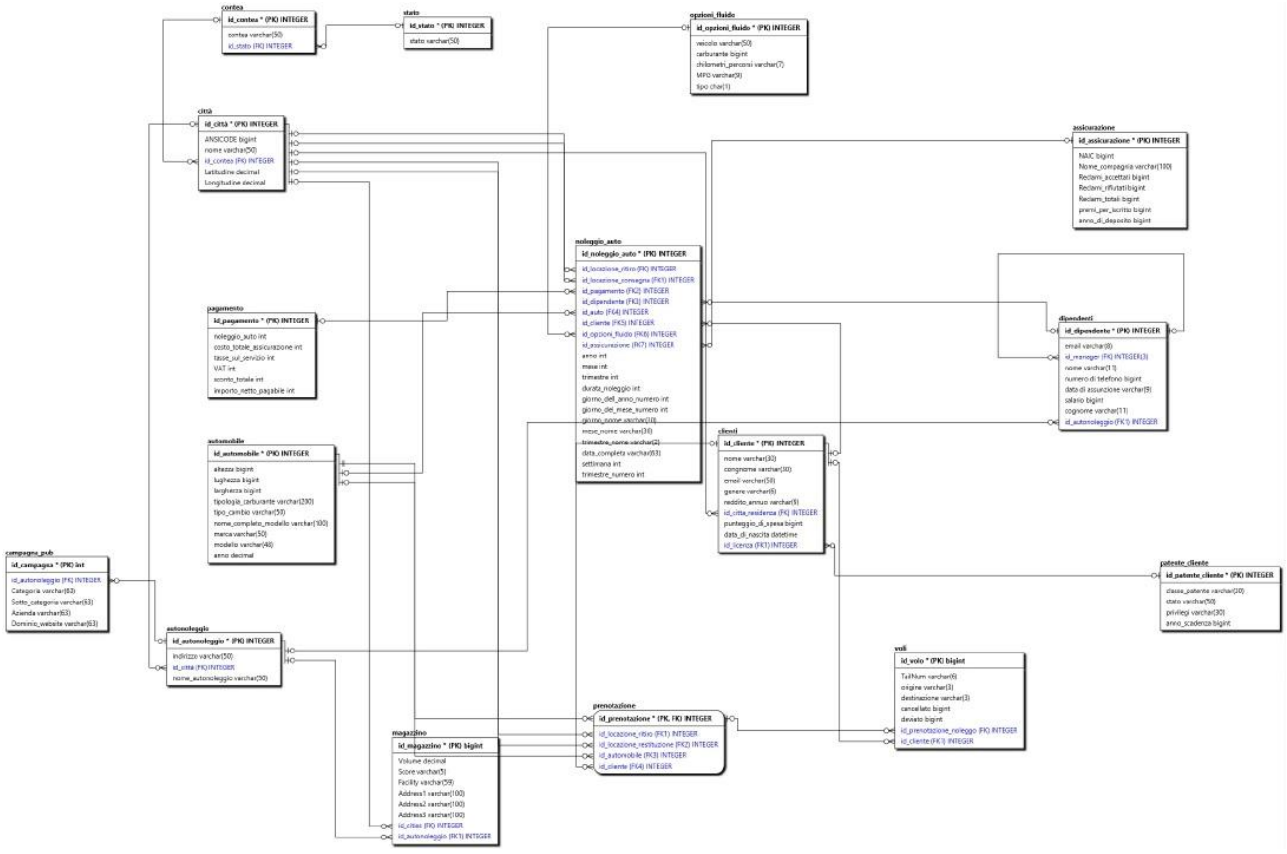
2.3 Integrazione

Le fonti dati che abbiamo utilizzato si integravano perfettamente tra di loro e fornivano informazioni di base abbastanza esaustive e complete. Per questo motivo oltre all'unione delle diverse informazioni in un unico DataBase non si è reso necessario integrare le sorgenti dati utilizzate con altre sorgenti.

2.4 Progettazione del livello riconciliato

Oltre a quanto fatto in precedenza, per la progettazione del livello riconciliato abbiamo attuato una fase di pulizia dei dati considerando solo le informazioni che ci sembravano rilevanti per poi andare ad alimentare il data warehouse. Tale livello è stato ottenuto attraverso una trasformazione in PDI che per ogni tabella in ingresso effettua alcuni controlli sull'unicità della chiave e sulla formattazione di alcuni attributi. Ovviamente le informazioni risultati verranno caricate nel DB relativo al livello riconciliato.

Di seguito riportiamo il modello ottenuto a seguito di questa fase.



3. Analisi dei Requisiti Utente

I soggetti di interesse che andranno ad usare il DW sono dei dipendenti statali o comunque analisti di un'azienda che hanno bisogno di fare analisi sui noleggi di automobili fatti nel territorio USA. Hanno bisogno di capire quali sono le zone in cui si noleggia maggiormente, cercando di capire quali sono gli autonoleggi più usati, le automobili più richieste e le compagnie assicurative che offrono costi minori.

Immaginiamo inoltre, che gli utenti siano interessati a fare analisi sui noleggi fatti su quali siano i luoghi di ritiro e consegna più richiesti.

Agli analisti interessa anche il reperimento degli autonoleggi più frequentati, e quelli che guadagnano maggiormente rispetto agli altri. Vogliono la possibilità di indagare sui meccanismi promozionali usati da questi negozi.

Si è interessati inoltre al numero medio di chilometri percorsi per un certo numero di noleggi identificati, per esempio, con il luogo di ritiro o con la data. Oppure si è interessati anche al carburante medio usato, oppure anche al MPG medio.

Infine, si vuole sapere da quali zone gli utenti noleggiano più frequentemente le automobili, e verso dove sono diretti.

Di seguito si riporta un esempio di glossario dei requisiti utente:

Fatto	***Possibili dimensioni***	***Possibili misure***
PRENOTAZIONE	Automobile, Cliente, Volo, Città	Numero delle prenotazioni
NOLEGGIO AUTO	Automobile, Assicurazione, Città, Dipendente, Opzioni Fluido, Pagamenti, Cliente	Carburante medio, chilometri medi, numero noleggi

A seguire un campione del carico di lavoro preliminare:

Fatto	***Interrogazione***
PRENOTAZIONE	Quantità di prenotazioni effettuate da un cliente. Quantità di prenotazioni effettuate per una specifica automobile.
NOLEGGIO AUTO	Quantità di noleggi effettuati complessivamente o per una specifica città/contea/stato. Costo assicurazioni per il noleggio di una particolare automobile. Età media dei clienti di uno specifico autonoleggio.

4. Pianificazione del DW

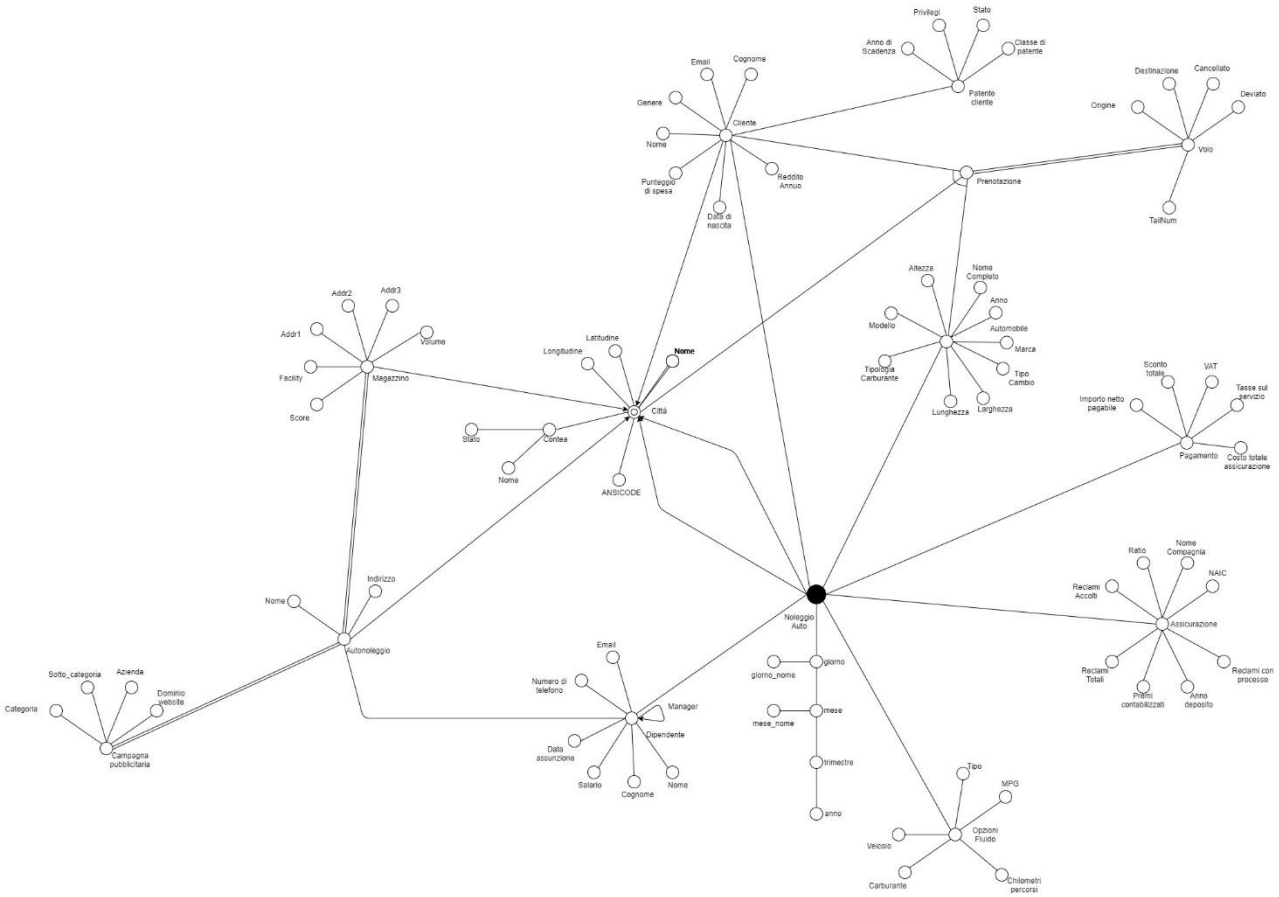
Vista l'analisi delle sorgenti dati operazionali abbiamo ritenuto che il fatto di maggiore interesse, per i soggetti identificati, sia il noleggio di una automobile.
La scelta è ricaduta sul noleggio poiché il numero di analisi (rilevanti per i soggetti di interesse) con le quali lavorare è nettamente maggiore rispetto a considerare come fatto le prenotazioni.

Inoltre, abbiamo deciso di adottare un approccio guidato dai dati per due motivi: da un lato, la presenza dello schema riconciliato ci ha permesso di guadagnare una conoscenza sufficientemente approfondita degli schemi operazionali così da rendere superfluo il ricorso ad un approccio guidato esclusivamente dai requisiti; dall'altro lato, la complessità dei dati operazionali non è tale da giustificare il ricorso ad un approccio misto.

Infine, le fasi di progettazione concettuale, logica e fisica del data mart sono state svolte con l'ausilio di ApricotDB, della suite Pentaho e come DBMS MySQL.

5. Progettazione Concettuale

Abbiamo riconosciuto come fatto di interesse il noleggio di un'automobile, e quindi a partire da questo fatto abbiamo definito l'albero degli attributi risultate.



Una prima scelta fatta è stata quella di cambiare la granularità del fatto eliminando la dimensione clienti. Abbiamo infatti ritenuto necessario che gli eventuali utenti di business orientati all'uso del DW non siano interessati a fare analisi sui clienti singoli che hanno fatto un noleggio, bensì su gruppi di utenti raggruppati o per il luogo in cui hanno ritirato l'automobile o per il luogo in cui l'hanno consegnata, oppure per l'autonoleggio a cui si sono riferiti.

Facendo questa scelta, come anticipato, la granularità del fatto è cambiata non considerando più i noleggi fatti da singoli clienti ma da gruppi di clienti.

Dopo di che procedendo dal fatto abbiamo potato il nodo relativo ai dipendenti, in quanto non di interesse per l'analisi da noi pensata, e per via della dipendenza funzionale tra dipendente e autonoleggio abbiamo collegato il fatto di interesse direttamente all'autonoleggio in cui è stato fatto il noleggio.

Sono state mantenute, inoltre, alcune informazioni relative ai dipendenti dell'autonoleggio ed alla loro cardinalità, questa scelta è stata fatta in quanto per un analista potrebbe risultare vantaggioso fare analisi considerando gli autonoleggi con un certo numero di dipendenti, per valutare in questo modo le prestazioni degli autonoleggi più grandi, con un numero di dipendenti maggiori. Un'altra possibile analisi interessante è quella relativa ad i noleggi effettuati in autonoleggi che hanno un determinato salario medio, anche in questo caso per valutare magari in relazione alla retribuzione dei dipendenti le prestazioni dei diversi autonoleggi.

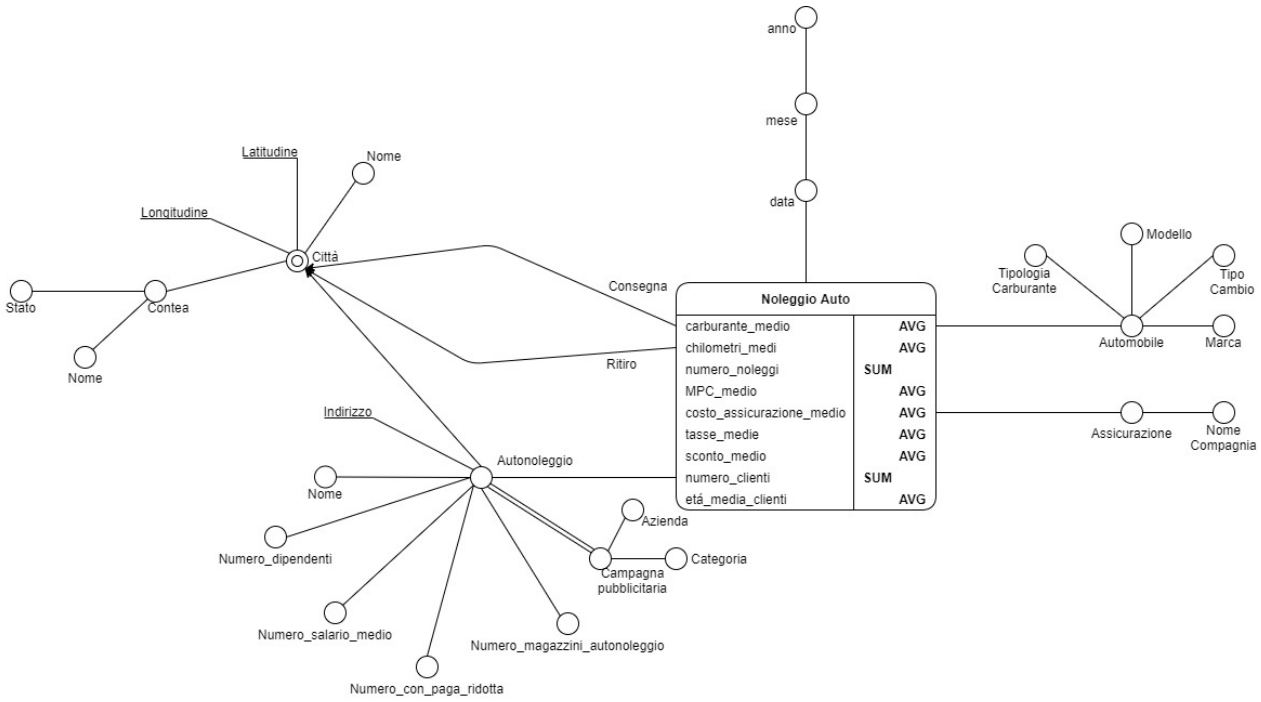
La gerarchia autonoleggio comprende anche la città in cui sono collocati gli autonoleggi, sono inoltre indicati con un arco multiplo tutte le compagna pubblicitarie fatte dall'autonoleggio stesso.

Oltre alla dimensione temporale le altre dimensioni sono:

- luogo di ritiro e di consegna dell'automobile, attraverso la gerarchia condivisa "città";
- l'automobile noleggiata, con informazioni sul modello e marca dell'auto;
- assicurazione, che contiene le informazioni sulla compagnia assicurativa che ha sottoscritto il contratto di noleggio.

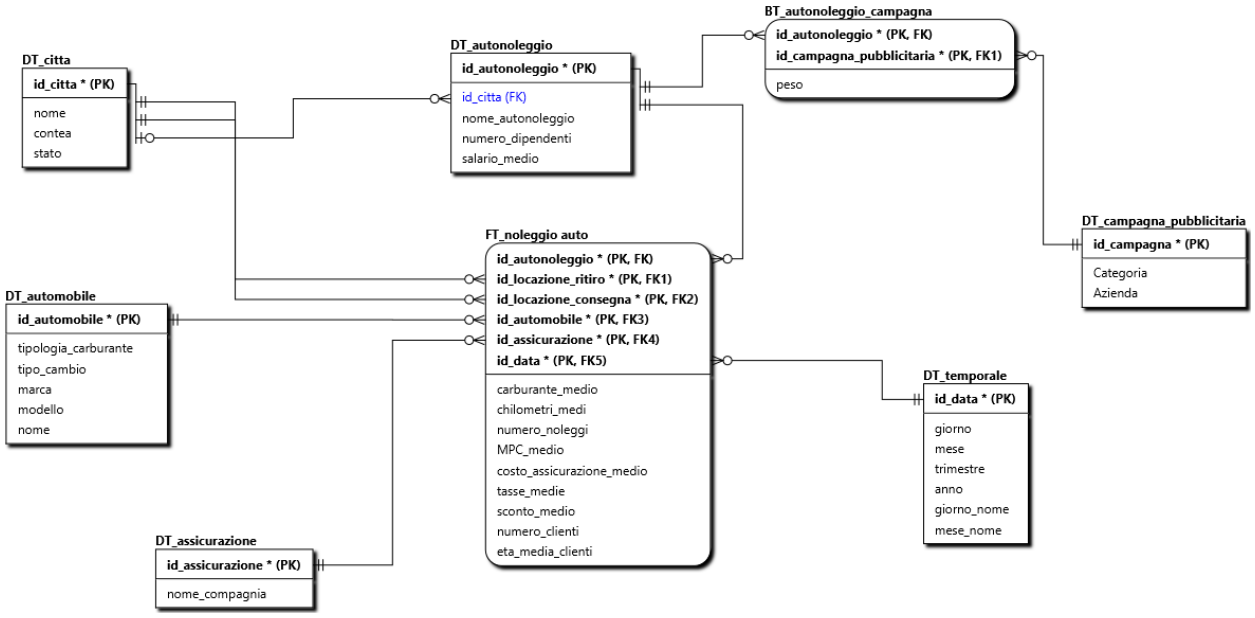
Le misure invece che richiedono di essere specificate, sono:

- età_media_clienti: Corrisponde alla somma delle età dei clienti che hanno effettuato il noleggio di un'automobile, firmando in una data, un determinato contratto assicurativo e noleggiando l'automobile in un autonoleggio. L'automobile, inoltre, è stata ritirata e consegnata (anche in luoghi distinti);
- numero_noleggi: Misura di supporto per aggregare per media.



6. Progettazione Logica

Lo star schema risultate dal DFM indicato è quello sotto riportato:

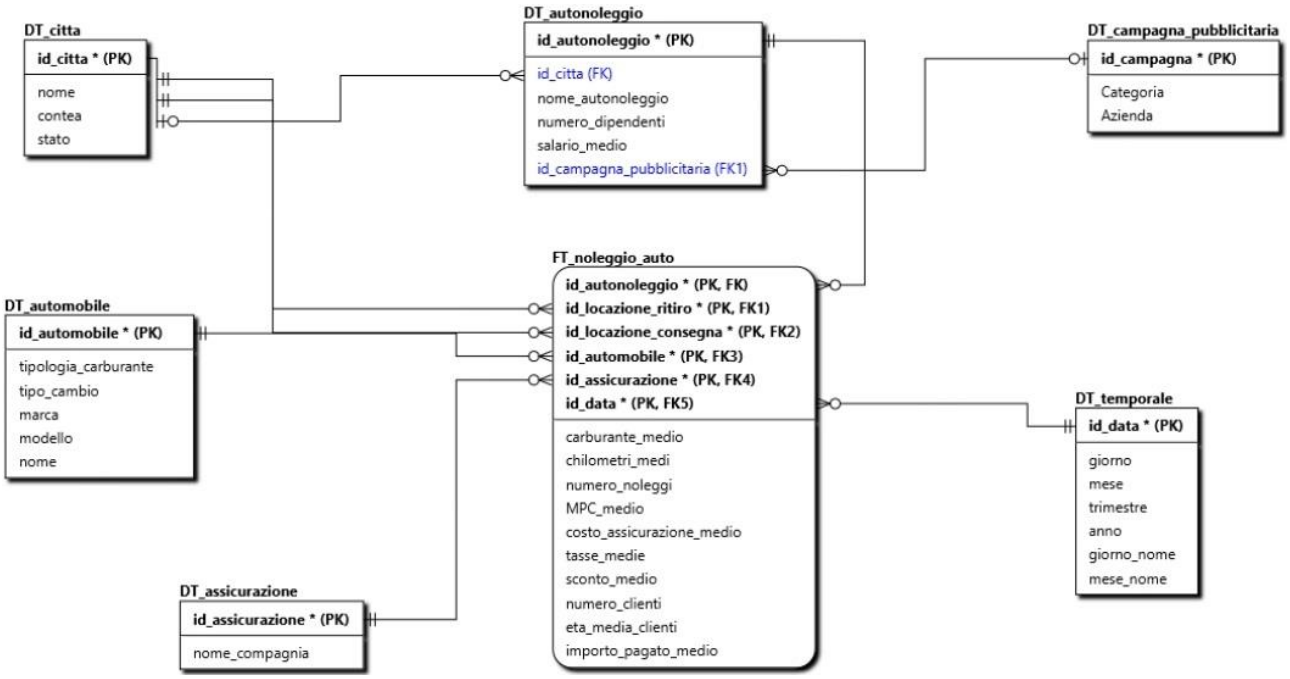


A livello logico, come si denota nella figura soprastante, si è optato per l'utilizzo di uno schema a stella identificando come dimensioni primarie, oltre quella temporale: assicurazione, automobile, città ed autonoleggio.

Gli scenari temporali che si dovranno supportare sono quelli di tipo "oggi per ieri" e di conseguenza abbiamo scelto di utilizzare le gerarchie dinamiche di tipo 1. Utilizzando quindi un classico schema a stella assicurandosi che a seguito dell'inserimento di valori aggiornati vengano sovrascritti quelli vecchi.

Per quanto riguarda l'arco multiplo relativo ad autonoleggio e campagna pubblicitaria, inizialmente, è stato gestito attraverso l'introduzione di una Bridge Table, che contiene le due chiavi relative agli attributi coinvolti, oltre che al peso. Dopo di che si è deciso di modificare la relazione molti-a-molti, che generava l'arco multiplo, con una relazione uno-a-molti così da mantenere la dimensione relativa a campagna pubblicitaria. Questa scelta, benché logicamente non perfetta, è stata fatta in quanto, per la creazione del cubo xml che sarà usato dal mondrian engine, non è possibile rappresentare relazioni molti a molti. Abbiamo optato quindi per una strategia che comprendesse per un autonoleggio una singola compagna pubblicitaria, intendiamo in questo modo l'ultima campagna pubblicitaria effettuata dall'autonoleggio.

Lo schema modificato sarà quindi:



Dopo aver fatto ciò abbiamo analizzato lo schema per valutare se fosse necessario fare snowflaking in qualche punto della gerarchia. Tuttavia non abbiamo rintracciato altri punti di snowflaking, qualsiasi possibile scelta avrebbe comportato infatti notevoli peggioramenti delle prestazioni, ciò è dovuto principalmente al fatto che le nostre dimensioni hanno delle gerarchie contenute.

Per quanto riguarda le possibili viste materializzate abbiamo ipotizzato che gli utenti del sistema siano fortemente interessati a fare analisi relative ad i modelli di automobili noleggiate in un particolare stato. Abbiamo inoltre assunto che è un tipo di analisti molto frequente.

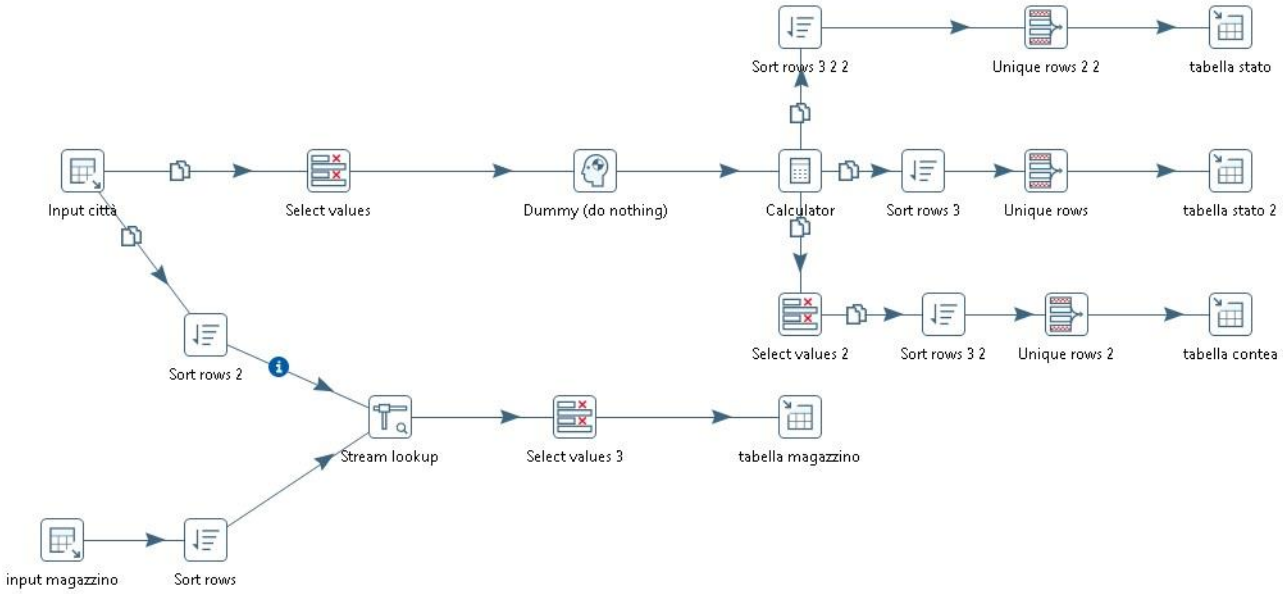
La prima assunzione ha comportato la scelta di materializzare la singola vista di interesse e non una ad un livello superiore del reticolo di roll-up, visto che comunque avrebbe avuto delle dimensioni maggiori e non sarebbe servita per fare analisi in modo più efficiente. La seconda assunzione fatta ci ha portato a non fare alcun tipo di snowflaking poiché non volevamo rallentare l'esecuzione delle query e la dimensione della fact table non è molto elevata (~10 MB contro i 100 MB della fact table originale).

7. Progettazione dell'Alimentazione

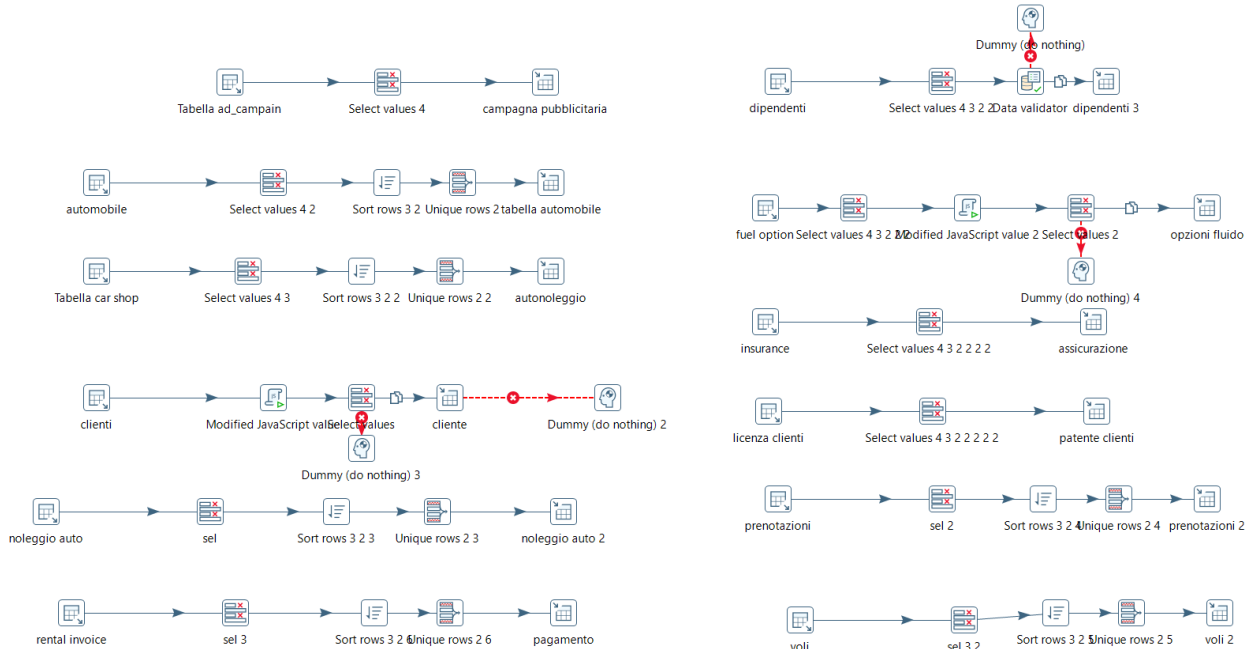
Prima di alimentare il livello riconciliato si è creato il DB operazionale attraverso delle semplici operazioni di estrazioni di informazioni da file e manipolazione degli stessi al fine di creare un DB consistente.

A partire da questo DB è stato creato il livello riconciliato normalizzando alcune entità (vedi paragrafo 2.2) e considerando solo le informazioni di interesse per l'analisi.

La normalizzazione è stata fatta con la seguente trasformazione:

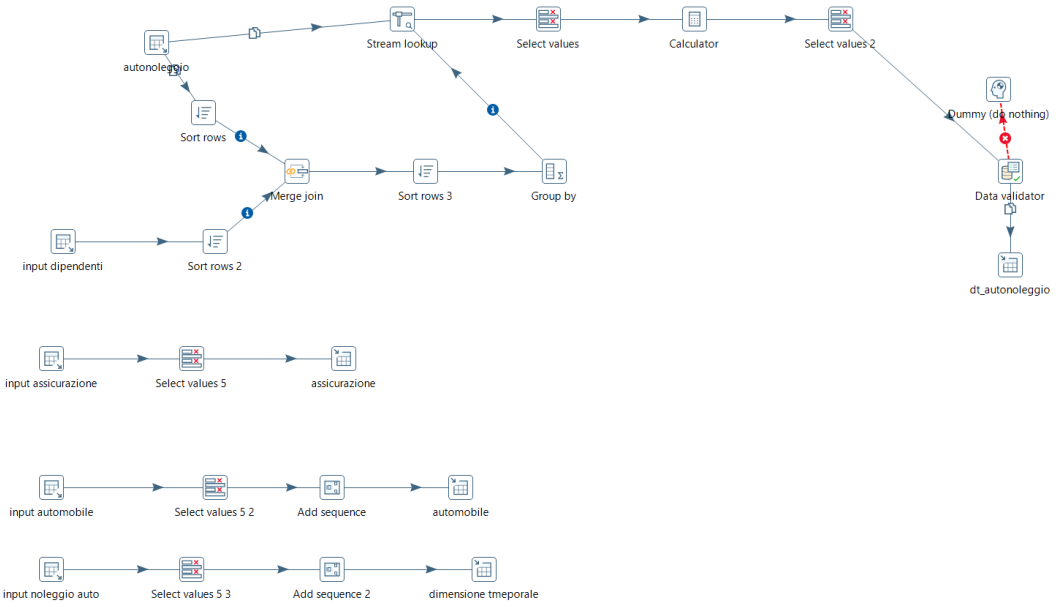


Per quanto riguarda il popolamento, si sono effettuate delle trasformazioni di passaggio dei dati ponendo particolare attenzione sulla verifica dell'unicità delle chiavi e sulla formattazione di alcuni attributi.



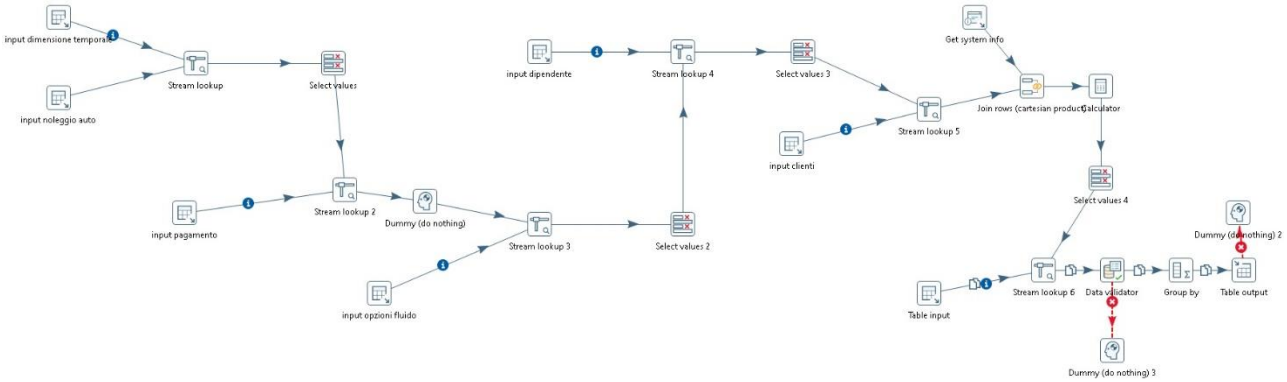
Per quanto riguarda la seconda fase, invece, dopo aver popolato il DB del livello riconciliato, si è passato a quello del DW partendo dalle dimensioni secondarie fino ad arrivare al fatto di interesse.

Il popolamento delle dimensioni primarie avviene prendendo le informazioni dal livello riconciliato ed inserendole nel DW. Per la creazione della dimension table autonoleggio è stato necessario effettuare una group by per riuscire a calcolare le informazioni cumulative relative ad i dipendenti. Inoltre, prima del caricamento nella tabella opportuna, è stata effettuata una validazione dei dati per eliminare eventuali inconsistenze.

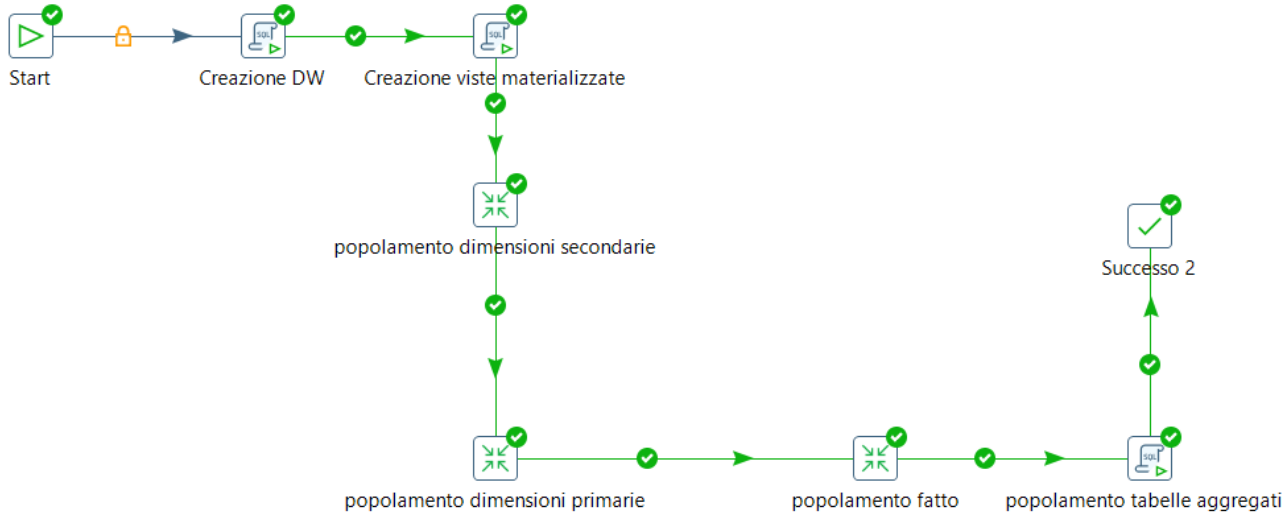


Il processo di caricamento delle fact table, è caratterizzato da un accesso al livello riconciliato per recuperare tutte le misure e i valori di chiave relativi alle diverse dimensioni.

Sono stati fatti inoltre più volte dei controlli per assicurarsi che i valori nella tabella noleggio auto siano uguali a quelli presenti nelle varie dimension table. Infine, prima del caricamento effettivo, è stata fatta una group by in modo da raggruppare esclusivamente per: locazione ritiro, locazione consegna, autonoleggio, automobile, assicurazione



Il tutto ovviamente è stato automatizzato attraverso l'introduzione di un singolo job che si occupa di effettuare tutte le operazioni riportate. Si occupa inoltre, attraverso passi SQL, di creare e popolare la vista materializzata scelta.



8. Progettazione fisica

L'aspetto principale da curare in questa fase è la scelta degli indici da costruire su fact e dimension table. Dall'analisi della manualistica di MySQL, su cui implementiamo il DW è emerso che, è dotato di un ottimizzatore basato su statistiche e quindi possiamo creare gli indici con tranquillità poiché, nel caso di implementazioni errate (predicato poco selettivo), sarà l'ottimizzatore stesso ad escludere l'indice dal piano di esecuzione.

Conseguenza dell'uso di chiavi surrogate per le dimensioni, si è optato, infine, per l'utilizzo di star index per l'evento primario. Questo perché, collegando le chiavi primarie delle dimension table con le corrispondenti chiavi importate nella fact table, con uno star index, si accelerano tutte le interrogazioni che coinvolgono la gerarchia modellata.