POLITECNICO DI TORINO

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE

TESI DI LAUREA MAGISTRALE

DATA QUALITY ASSESSMENT: IL VALORE DELL'INFORMAZIONE NEL PROCESSO DI DIGITAL TRANSFORMATION DI UN DEALER.

Relatore

Tania Cerquitelli

Candidato

Viola Burresi

Marzo 2020

RINGRAZIAMENTI

Il ringraziamento più importante che io possa fare è a Gianluca, capace di capirmi senza chiedere, di sostenermi sempre e di regalarmi un sorriso con la sua sola presenza.

Grazie a tutta la mia famiglia, ma in particolare a mia Mamma, che c'è sempre stata nei momenti di sconforto ed i cui consigli sono stati ascoltati più di quanto lei creda.

Grazie a Ginevra, Malvina, Gabriella e Valeria, che mi hanno fatto capire che la vera amicizia non è vedersi tutti i giorni, ma di ritrovarsi e vedere che niente è cambiato nonostante la distanza.

Vorrei ringraziare Gian Luca De Michelis, per essere un ottimo tutor e per aver creduto in me fin dall'inizio di questo percorso.

Infine, ringrazio le persone conosciute in questi anni e i colleghi di lavoro, con i quali ho passato e passo momenti bellissimi.

SOMMARIO

INTRODUZIONE

Questo lavoro nasce in seguito all'esperienza maturata nel periodo di stage svolto presso l'azienda Engineering Ingegneria Informatica Spa in affiancamento e supporto al lavoro del Project Manager Tecnico dell'ufficio Engineering Automotive Business Unit, per progetti focalizzati sulla trasformazione digitale dei gruppi Dealer. Durante questo periodo ho avuto modo di approfondire tematiche riguardanti la gestione del software e sul processo di sviluppo del sistema. L'attività che mi ha vista più coinvolta è stata la gestione e monitoraggio delle evoluzioni delle strutture dati a supporto dei moduli della Digital Dealer Platform, una soluzione creata da Engineering Ingegneria Informatica spa per la digitalizzazione dei processi di business dei Dealer.

Durante questo periodo è stata svolta un'analisi critica dei problemi di data integration riscontrati fino ad ora per differenti progetti, che ha portato all'individuazione di criticità durante questa fase di integrazione dati, nello specifico nella fase di analisi della qualità dei dati, molto spesso sottostimata in fase iniziale di offerta al cliente.

I grandi gruppi concessionari stanno vivendo oggigiorno un periodo di forte cambiamento e proprio in questo contesto di evoluzione, stanno attualmente emergendo, per i Dealer, necessità di utilizzare servizi in grado di semplificare ed automatizzare i processi operativi ed in generale, di ottimizzare le funzioni di vendita di veicoli, accessori e servizi connessi. La risposta a queste esigenze passa attraverso la gestione, il trattamento e l'analisi di una grande quantità di dati proveniente da fonti diverse, non integrate tra loro, come i sistemi interni di proprietà (OMS, CRM, ecc.), i sistemi forniti dalle case madri relativi a tutti i brand di cui il Dealer ha il mandato o sistemi forniti da provider di servizi e altri dati esterni (es: assicurazioni, banche, siti usato, ecc.).

È chiaro, quindi, come in questo contesto i concessionari necessitino sempre più di strumenti di gestione che permettano di governare, controllare e semplificare i processi tipici della concessionaria. La Digital Dealer Platform si inserisce in questo contesto come una proposta di piattaforma adibita all'integrazione di tutti i dati interni ed esterni all'ecosistema aziendale, con lo scopo di renderli fruibili per il proprio business.

Una delle problematiche tipiche di una integrazione di dati provenienti da molteplici fonti riguarda proprio la definizione di regole per stabilire la qualità del dato all'interno del sistema, in modo da renderlo fruibile per le attività di gestione e di analisi.

L'obiettivo della tesi è l'individuazione e definizione di indicatori della qualità del dato, proveniente da un processo di integrazione e determinati in seguito ad un'analisi del database. Questi indicatori saranno utilizzati per monitorare nel tempo l'evoluzione della qualità del database. Verrà presentata anche una possibile dashboard in modo da poter meglio visualizzare il risultato ottenuto.

Nella prima parte dell'elaborato viene offerta una descrizione del contesto in cui opera il cliente e viene fatta una descrizione dell'attuale scenario del settore automotive. Viene mostrato come la trasformazione digitale stia influenzando il ruolo del concessionario nell'attuale scenario competitivo.

Nella seconda parte vengono illustrati i concetti teorici che stanno alla base dei processi ETL e i concetti di data quality e misura della qualità del dato.

Nella terza parte viene descritto il processo di analisi effettuata, nella conclusione a fornire degli indicatori della qualità del dato validi nel contesto di business descritto.

STRUTTURA DELLA TESI

Il primo capitolo della tesi fornisce le conoscenze di contesto necessarie alla comprensione delle dinamiche del settore automotive. Questo è alla base della comprensione della rivoluzione radicale che l'industry si appresta a subire e, di conseguenza, del perché il tema della digital transformation sia essenziale per gli attori della distribuzione, focalizzando l'attenzione sui Dealer e come stia cambiando il loro ruolo all'interno di questo nuovo contesto.

Nel secondo capitolo vengono illustrati i concetti teorici fondamentali dei processi di data integration e data quality, cosa significhi determinare il livello di qualità dei dati e come sia difficile darne una definizione univoca.

Il terzo capitolo approfondisce e descrive la struttura della piattaforma digitale DDP, analizzando in particolare le fonti di dati dei Dealer ed il processo di integrazione di questi nel repository centrale di questa soluzione, fornendo una panoramica sui principali sistemi operazionali di un Dealer.

L'ultima parte della tesi tratta nel dettaglio l'analisi svolta durante il periodo di tirocinio ed inserimento all'interno dell'azienda Engineering. In questo capitolo verrà presentata una descrizione del database in modo da definire il contesto di analisi ed individuare le entità critiche oggetto di studio. L'analisi si baserà su una serie di interrogazioni al database e successiva rielaborazione dei dati restituiti, in modo da poter definire delle metriche di misura della qualità del dato, nel contesto dei processi di un Dealer. Verrà poi presentata una dashboard per la visualizzazione dei risultati ottenuti.

LISTA DELLE ABBREVIAZIONI E DELLA TERMINOLOGIA UTILIZZATA

DEALER	Concessionaria automobilistica; nel contesto della tesi, si userà il termine Dealer per indicare indistintamente concessionarie mono-marca o multi-brand.
DMS	Dealer Management System – Sistema software di gestione per concessionari d'auto.
OEM	Original Equipment Manufacturer – Si riferisce ad un'azienda che realizza un'apparecchiatura che verrà poi installata in un prodotto finito, sul quale il costruttore finale appone il proprio marchio. Nel contesto della tesi, il termine OEM è usato nell'accezione di Casa Madre.
CRM	Customer Relationship Management: si tratta di una categoria d software per la gestione delle relazioni con il cliente
CMS	Content Management System - È uno strumento software, tipicamente installato su un server web, con un front end principale rivolto al pubblico (consumatore dei contenuti) ed una sezione di backend per l'inserimento e la gestione dei contenuti stessi.
Lead	Potenziale acquirente di un dato prodotto o servizio. Si genera un lead quando, attraverso un'iniziativa di marketing, un'impresa ottiene dall'utente informazioni utili a stabilire un contatto commerciale.
SFTP	Protocollo di rete
HTTPS	HyperText Transfer Protocol over Secure Socket - Protocollo per la comunicazione sicura attraverso una rete di computer utilizzato su Internet.
WS	Web Service - Sistema software progettato per supportare l'interoperabilità tra diversi elaboratori su una medesima rete oppure in un contesto distribuito.
API	Application Programming Interface - Interfaccia di programmazione delle applicazioni: sono set di definizioni e protocolli con i quali vengono realizzati e integrati software applicativi.
RDBMS	Relational Database Management System - Sistema per la gestione di basi di dati relazionali: indica un database management system basato sul modello relazionale
Dashboard	Letteralmente "cruscotto", è la visualizzazione di informazioni che aiuta a monitorare eventi e attività tramite tabelle e grafici
Business Intelligence	Insieme di processi aziendali per raccogliere dati ed analizzare informazioni strategiche

CAPITOLO 1 – LA TRASFORMAZIONE DIGITALE NEL SETTORE AUTOMOTIVE E FOCUS SULLA FIGURA DEL DEALER

1.1 IL SETTORE AUTOMOTIVE

Il settore Automotive è oggigiorno uno dei più importanti settori a livello mondiale. Si stima, infatti, che la vendita di auto mondiale abbia raggiunto la cifra di circa 95 milioni a fine 2018 (OICA, World Motor Vehicle Sales). Come si può evincere dalla figura 1, l'Europa rappresenta circa un quinto delle immatricolazioni di autoveicoli nel mondo e in Italia solo questo settore dell'autoveicolo vale il 10% del Pil nazionale considerando tutto l'indotto, diretto ed indiretto.

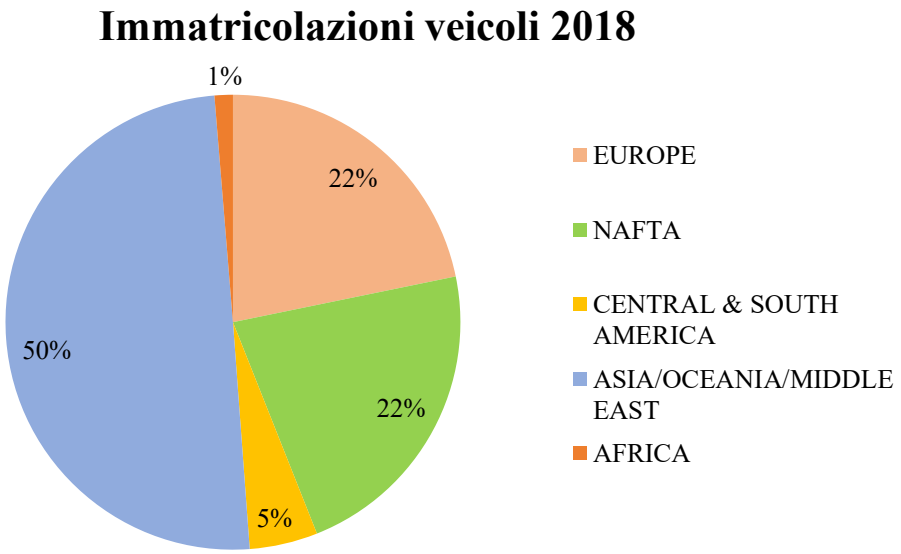

Figura 1 - Immatricolazioni veicoli nell'anno 2018 [OICA]

Figura 2 - Immatricolazioni autoveicoli nel 2018 (Totale: autovetture, VCL, autocarri, autobus) [ANFIA]

Il comparto Automotive è caratterizzato da un alto dislocamento fisico della catena di produzione: basti pensare che tocca cinque continenti, ed il consumo avviene su scala globale. Oltre ad un dislocamento fisico si ha, inoltre, un dislocamento produttivo: ogni euro di valore in fase di assemblaggio del prodotto finito genera circa 10 euro di produzione componentistica.

Il grafico raffigurato in figura 3 rappresenta i dati presentati nel 2018 ed illustra le statistiche dei veicoli venduti dal 2007 al 2018 nei principali mercati mondiali. Negli ultimi 10 anni la produzione mondiale è aumentata del 33 per cento, un incremento che vale circa 23,5 milioni di autoveicoli. Le aree con produzioni inferiori a quelle del 2007 sono state L'Africa (-7 per cento) e l'Unione Europea (-10 per cento). La produzione di autoveicoli in Unione Europea si è ridotta rispetto a dieci anni fa (toccava quasi il 27 per cento della produzione mondiale nel 2007, mentre oggi solo il 19,7), magli ultimi due anni sono stati comunque migliori del 2016.

I primi produttori al mondo di auto si confermano la Cina, gli Stati Uniti e il Giappone, che da soli producono oltre il 40 per cento dei veicoli totali.

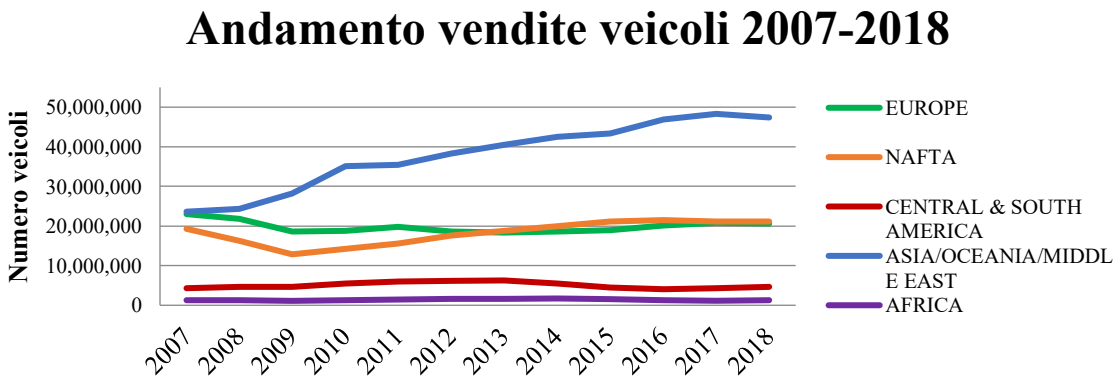

Figura 3 - Andamento vendite veicoli 2007-2018 [ANFIA]

1.2 LA VALUE CHAIN DEL SETTORE AUTOMOTIVE

Il settore automobilistico è stimato tra le industrie manifatturiere più grandi in tutto il mondo e si tratta di un settore considerato come altamente concentrato, poiché le aziende produttrici di veicoli sono un numero limitato e rappresentano una quota significativa della produzione. Lo scenario attuale ha visto negli ultimi anni l'aumento di partnership e fusioni, come ad esempio il più recente accordo tra Fiat Chrysler Automobiles ed il Gruppo PSA che porterebbe alla nascita di un gruppo automobilistico in grado di vendere quasi 9 milioni di auto e di piazzarsi al quarto posto nella classifica globale dei costruttori. Entro il 2030 si prevede però una flessione della quota di ricavi derivanti dalle vendite di auto tradizionali, al contrario della manutenzione post-vendita, delle vendite di pezzi di ricambio e dei servizi connessi all'uso di auto da cui si prevede afferiranno la maggior parte dei ricavi.

Recenti studi hanno dimostrato che per ogni moneta investita nell'industria automobilistica, questa fa aumentare il PIL del Paese di circa tre volte, ciò significa che si hanno alti profitti e benefici dove questo settore è maggiormente sviluppato.

In effetti, l'industria automobilistica è una delle principali fonti di sviluppo per un paese: Stati Uniti, Giappone, Germania e Corea del Sud sono un esempio di "super industria". Il fenomeno della globalizzazione ha portato vantaggi per gli operatori del settore automobilistico, grazie maggiormente alla standardizzazione dei modelli sia in diversi paesi che mercati.

Questa standardizzazione comporta però non pochi problemi: infatti, non tutti i veicoli sono adatti per essere venduti negli stessi customer segment e questo può portare ad una riduzione dei profitti. I produttori dei veicoli devono adeguare l'offerta alle diverse esigenze dettate dai diversi mercati che rispecchiano i bisogni dei clienti finali.

La rete di distribuzione automobilistica svolge un ruolo fondamentale aiutando i consumatori nella scelta giusta. Considerando le fasi industriali e quelle di distribuzione, la catena di approvvigionamento globale contribuisce in Italia per il 5% del PIL nazionale.

L'industria di questo settore, per ottimizzare i propri processi produttivi si basa sull'efficienza nell'integrazione delle differenti fasi della produzione e sull'esternalizzazione di alcune di queste fasi. Vengono sfruttate le economie di scala, l'ottimizzazione dei costi e gli alti standard dei livelli di produzione tramite accordi specifici con partner altamente qualificati.

Oggigiorno si ha un cambiamento nella percezione dei veicoli da parte dei consumatori: i benefici del "car-as-a-service" sono più allettanti del design stesso. Le grandi imprese di assemblaggio dei componenti hanno visto da sempre eneternalizzata la fase di produzione dei vari componenti, ma le attività core del settore restano comunque integrate verticalmente all'interno dell'azienda.

I fornitori sono geograficamente dispersi e recentemente le imprese hanno iniziato a collaborare con altri partner, creando una catena del valore unica che garantisse così di ottenere una massimizzazione dei vantaggi competitivi. Questa situazione è dettata dalla possibilità di beneficiare, da parte delle grandi aziende automobilistiche, dei bassi costi di produzione e delle materie prime, dei paesi in via di sviluppo.

Le nuove tecnologie stanno sfidando il settore e tutti gli attori della catena del valore (guida autonoma, elettrificazione e connettività). Tutti questi fattori stanno facendo pressione sulle aziende coinvolte in termini di maggiore flessibilità nella catena di fornitura e di riduzione dei costi di produzione. Con il cambiamento del comportamento dei clienti e una maggiore richiesta di personalizzazione, la flessibilità diventa fondamentale nella produzione automobilistica. Inoltre, l'integrazione delle nuove tecnologie nei veicoli richiede processi sempre aggiornati e nuovi. Si è creata la necessità di implementare nuovi modi di lavorare per poter abbracciare pienamente i cambiamenti e questa nuova complessità nel settore automobilistico.

La catena del valore include tutte le attività e i processi aziendali che generano valore aggiunto, dalla produzione del prodotto fisico finale e dal suo valore economico. Ogni fase della catena è altamente correlata con la precedente e la successiva: la relazione tra di loro è di solito di lunga durata. La catena del valore del settore automobilistico può essere suddivisa in due macro-aree: la catena di approvvigionamento del prodotto fisico e i servizi connessi alla sua produzione. Si ha, inoltre, un'integrazione verticale tra le diverse fasi: esiste una rete di fornitori sia a monte che a valle.

Gli attori principali di questa catena del valore sono: i fornitori (primo-secondo-terzo livello), le case automobilistiche (OEM) e la distribuzione (Dealer).

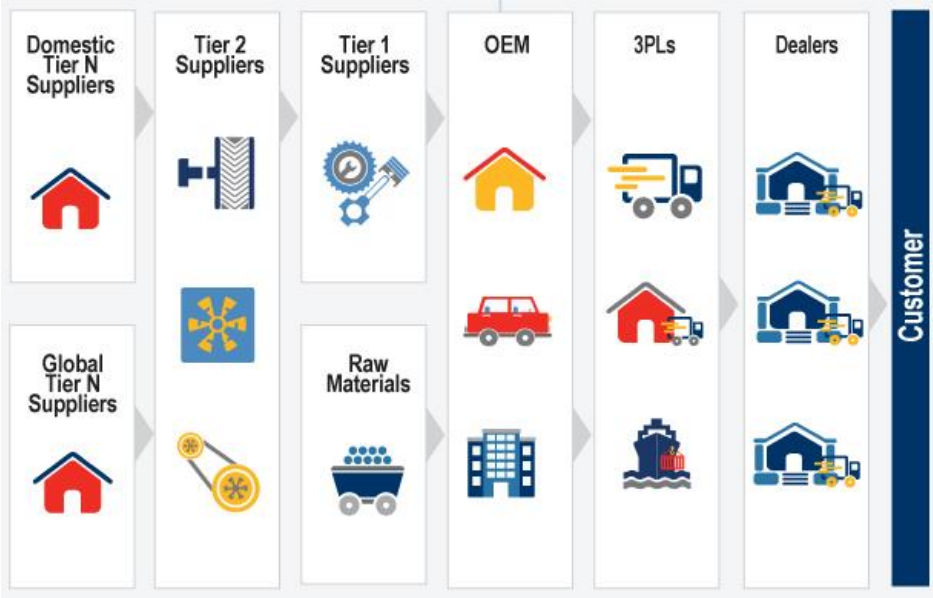

Figura 4 - Rappresentazione degli attori della catena del valore nel settore automotive

1.3 LA DISTRIBUZIONE AUTOMOTIVE IN ITALIA

Così come lo scenario europeo, anche il mercato italiano sta procedendo nello stesso modo: infatti, nel contesto italiano, l'industria è principalmente formata da imprese a conduzione familiare collocate nei centri urbani tramite punti vendita.

Con l'aumento delle concessionarie, è aumentato il bisogno di vendere ulteriori servizi, oltre al prodotto automobilistico, come ad esempio assicurazioni, estensioni di garanzia, leasing e altri servizi.

I Dealer appresentano oggi solo il 5% degli imprenditori nella distribuzione auto ma valgono circa un quarto dell'intero mercato, per un valore salito a oltre 15 miliardi di euro. I 50 top concessionari italiani stanno accelerando nonostante un 2018 in recessione e segnano una crescita del 10% sul 2017, portando la progressione del fatturato medio nell'ultimo decennio a +54%.

Questo è lo scenario presentato da Quintegia[1] nell'ambito dell'Annual Meeting - Top Dealer Network, l'incontro a porte chiuse della community dei più grandi dealer italiani. Il risultato esposto sopra è l'effetto di un processo senza precedenti di fusioni e acquisizioni, che solo negli ultimi 18 mesi ha visto i primi 10 gruppi per fatturato in Italia mettere a segno 19 nuove acquisizioni. Tra questi, il gruppo Eurocar con 5 acquisizioni, Autotorino con 2 acquisti e una fusione (con Autostar), Penske (1 acquisizione) e Bossoni con 3.

Tra i marchi più rappresentati, primeggiano Fiat e Alfa Romeo - presenti nel 38% dei casi - seguiti dai premium Jeep (34%), Bmw, Mercedes e Volkswagen (32%), Volvo (28%) e Audi (26%), con Skoda, Toyota e Nissan sopra il 20%.
I 10 top dealer, rileva l'analisi Quintegia, hanno sede prevalentemente nel Nord del Paese e realizzano mediamente 600 milioni di euro con quasi il 10% delle auto nuove vendute in Italia. Un valore del fatturato che nell'ultimo decennio ha registrato un incremento del 67%, a riprova del consolidamento avviato nella rete distributiva italiana che nello stesso periodo ha visto chiudere le saracinesche al 60% dei propri imprenditori.

Gli anni 2015 e 2016 hanno visto un decremento delle piccole imprese concessionarie. La causa principale di questa diminuzione è dovuta agli importanti cambiamenti che tutto il settore sta vivendo e alle novità che si stanno affacciando in ambito mobility. Questi cambiamenti hanno portato all'incremento di fusioni tra aziende in modo che l'aumento della loro dimensione nel mercato della distribuzione potesse essere un elemento importante in modo da garantire margini operativi migliori e un investimento in tecnologie innovative e sulla trasformazione digitale dei loro sistemi aziendali. L'avvento di queste nuove situazioni, all'interno del settore automobilistico, e la riduzione degli attori coinvolti ha permesso un aumento delle vendite per ogni Dealer e una maggiore profittabilità. Anche il modo di gestione delle marche, all'interno delle concessionarie, è cambiato

[1] Dal 2003 Quintegia è un punto di riferimento per l'ecosistema di business del settore automotive mediante una serie di attività rivolte ai principali attori della filiera distributiva.

sostanzialmente: se prima un Dealer gestiva un solo marchio, adesso un solo Dealer ne arriva e gestirne più di uno.

Un aspetto importante e che evidenzia risultati rilevanti è la dimensione del concessionario. Il 2012, come per altri settori, è stato un anno molto impegnativo per questo mercato, i concessionari di piccole e medie dimensioni hanno visto diminuire i propri profitti dell'1%, cosa che non si è verificata per i gruppi di grandi dimensioni. Le difficoltà per i piccoli concessionari sono aumentate a causa anche delle pressioni e degli alti tassi di interesse applicati dalle banche in questo periodo di grande difficoltà.

In confronto allo scenario europeo, tuttavia, i dealer italiani presentano dimensioni minori e nessuno è presente su tutto il suolo italiano, infatti è presente una segmentazione nord, centro e sud delle aree servite. Il gruppo concessionario più grande in Italia risulta Autotorino con una quota di mercato di circa del 2%. Questa quota va tuttavia confrontata con le più alte quote di UK, Germania e Francia rispettivamente del 28%, 20% e 14%. Gli ultimi anni hanno visto l'entrata nello scenario italiano delle grandi compagnie straniere, come ad esempio l'americana Penske e la tedesca Porsche Holding, che procedono con l'acquisizione di alcuni Dealer presenti nel nord Italia.

Il processo di aggregazione dei piccoli concessionari negli anni è stato favorito anche dalla possibilità di ottimizzare la gestione dei grandi investimenti e la riduzione dei rischi connessi all'attività. La distribuzione dei concessionari in Italia, in generale, rispecchia la suddivisione territoriale presenti anche negli altri settori manifatturieri, infatti il 55% dei concessionari si trova al nord, il 20% al sud ed il 25% nel centro dell'Italia. Il fatturato dei concessionari è principalmente composto per il 47% dalla vendita del nuovo; la restante percentuale è suddivisa in vendita dell'usato, dei pezzi di ricambio e assistenza.

Tutto il settore automobilistico sta affrontando profondi cambiamenti e questo si riflette su una nuova redistribuzione della catena del valore. La connettività e i nuovi sviluppi

tecnologici impongono ai concessionari di avere maggiori contatti diretti con gli utenti finali. Si prevede che entro il 2030 una differente distribuzione del valore aggiunto in questo settore (espresso in miliardi di dollari): le aziende fornitori impatteranno per il 5%, il 26 % deriverà dalla vendita dei veicoli nuovi, i servizi post vendita e le assicurazioni e finanziamenti avranno un peso del 20% e la restante parte sarà in mano ai fornitori di tecnologie e servizi di mobilità.

1.4 LA DIGITAL TRANSFORMATION ED IL NUOVO RUOLO DEI DEALERS

Si sente parlare molto di Digital Transformation, ma spesso non è chiaro cosa si intende. In genere, la trasformazione digitale è intesa come l'utilizzo di nuove tecnologie e canali di informazione per migliorare i processi interni e la gestione dei flussi informativi, generare un contatto più diretto e personalizzato con il cliente ed incrementare le performance delle aziende.

Tuttavia, soffermarsi solo sull'aspetto tecnologico è limitante alla comprensione di questo movimento che sta cambiando il mondo. La Digital Transformation non è solo l'effetto dell'investimento nelle nuove tecnologie abilitanti, ma un processo guidato da una strategia di rivisitazione e ridisegno dei modi di lavorare, dei modelli di business, dei modi di comunicare, i nuovi paradigmi digitali da abilitare, strettamente legati all'utilizzo del dato.

Il MIT di Boston la descrive usando tre dimensioni fondamentali: trasformazione della customer experience, trasformazione dei processi operativi, trasformazione dei modelli di business.

Il settore Automotive è, oggi più che mai, un mercato altamente complesso e concorrenziale, in cui la figura del cliente sta assumendo un ruolo decisamente

importante, diventando sempre più esigente e attento ai più piccoli particolari: stiamo, infatti, assistendo ad un cambiamento dei modelli con cui i Dealer si relazionano con i clienti e con essi le modalità di fruizione di servizi e contenuti.

Il fenomeno della digitalizzazione nell'industria automobilistica ha visto ridurre oggigiorno il bisogno di contato diretto con il concessionario. I clienti sono in grado di accedere alle informazioni sugli autoveicoli tramite i servizi presenti sulle App e sui portali istituzionali delle case automobilistiche, oltre alla possibilità di utilizzare i social network: viene stimato che la numerosità di clienti che effettivamente visita la concessionaria è diminuito, dato dal fatto che questa attività è considerata utile solo in caso di acquisto. Questo cambiamento ha impattato fortemente sul ruolo tradizionale della concessionaria. A livello nazionale, molti Dealer non sono riusciti ancora a strutturare un canale che permetta un'approfondita esperienza di interazione con il cliente digitale e solo pochi hanno intrapreso un percorso di trasformazione digitale dei processi di business volti a coinvolgere i potenziali customer tramite servizi online. Un problema derivante dall'utilizzo dei canali web sta nel fatto che i clienti, quando arrivano nella concessionaria, dispongono già di tutte le informazioni precise riguardanti i prezzi e le condizioni di acquisto, ciò comporta una riduzione del potere di contrattazione da parte del Dealer e una diminuzione della marginalità.

Lo scenario in cui operano i Dealer e i rivenditori sta cambiando con il consolidamento di nuovi canali di vendita e l'emersione di nuove forme di mobilità. La digitalizzazione è la forza motrice dietro questo cambio strutturale ed influenza ogni aspetto dello spirito imprenditoriale. L'onda della Digital Transformation sta spostando il focus principale del business: il Dealer non è più solo un fornitore di prodotto in senso stretto ma sta diventando un fornitore di servizio in senso lato.

Questa trasformazione digitale sta quindi introducendo nuovi modelli di business, in cui i servizi ricopriranno un ruolo maggiore rispetto ai prodotti iniziando ad introdurre il modello

di Product-as-a-Service, ovvero il concetto di prodotti che tramite i dati generati da oggetti interconnessi vengono fruiti come se fossero dei servizi.

D'altronde, nonostante stia cambiando l'esperienza di vendita dei clienti, il concessionario rimane comunque un punto cruciale nel percorso di decisione del cliente, infatti il rapporto che si crea in seguito ad un contatto diretto tra il rivenditore ed il cliente ricopre un ruolo centrale, specialmente in quei casi di informazioni contrastanti o di decisione sulla scelta del brand o modello di veicolo. Dopo la vendita, gli OEM continuano a richiedere il supporto dei Dealer per servire come brand locali e per fornire ai clienti con veicoli di alta qualità un servizio di manutenzione.

In questo contesto la Digital Transformation si pone come un percorso che ha come obiettivo creare soluzioni innovative per assistere e aiutare i Dealer a svolgere la propria missione in modo consistente e coerente, dotandoli di soluzioni in grado di aumentare le capacità di percezione dei segnali di contesto, di elaborazione delle informazioni raccolte e di interpretazione e automazione delle azioni da compiere.

Questo cambiamento non solo impatta i modelli di business, ma le Dealership devono affrontare anche una trasformazione dei processi operativi. Oggigiorno i processi di acquisto e di manutenzione dell'auto richiedono un livello crescente di integrazione della dimensione fisica e digitale al fine di creare un'esperienza omnicanale e più personalizzata su ogni cliente. Questo cambiamento consente la raccolta di un'ingente quantità di dati che, se utilizzati nel modo corretto, possono permettere alle Dealership di ottenere un vantaggio competitivo in termini di CRM, Customer Experience Loyalty, aiutando davvero a costruire una relazione focalizzata sul lungo periodo tra la Dealership ed il Cliente finale.

Un sistema CRM ha la funzione di potenziare il business e di permettere una gestione integrata di tutti i dati derivanti dai sistemi DMS e di terze parti che interagiscono con la

concessionaria. La raccolta dei dati inizia nel momento in cui il cliente genera un lead online oppure varca la soglia dello showroom, e continua durante tutto il customer journey, partendo dalle operazioni che riguardano il processo di gestione della trattativa e di acquisto dell'autovettura, fino ad arrivare alla gestione post vendita.

È chiaro, quindi, come in questa nuova era digitale i concessionari necessitino sempre più di strumenti di gestione che permettano di governare, controllare e semplificare i processi tipici della concessionaria.

CAPITOLO 2 – DATA INTEGRATION E DATA QUALITY:

CONCETTI FONDAMENTALI

Tra i sistemi di supporto alle decisioni, i sistemi di data warehousing sono probabilmente quelli su cui negli ultimi anni si è maggiormente focalizzata l'attenzione sia nel mondo accademico sia in quello industriale. Una definizione informale di data warehousing è stata data da Golfarelli e Rizzi:

"Il Data warehousing è una collezione di metodi, tecnologie e strumenti di ausilio al cosiddetto "lavoratore della conoscenza" per condurre analisi dei dati finalizzate all'attuazione di processi decisionali ed al miglioramento del patrimonio informativo."

Al centro del processo vi è il data warehouse, un contenitore di dati che diventa garante dei requisiti esposti. William Inmon ne diede una definizione nel 1996:

"Un Data warehouse (DW) è una collezione di dati di supporto per il processo decisionale che presenta le seguenti caratteristiche: è orientata ai soggetti di interesse; è integrata e consistente; è rappresentativa dell'evoluzione temporale e non volatile."

Il DW è orientato ai soggetti in quanto in quanto si incentra sui concetti di interesse dell'azienda, quali clienti, i prodotti, le vendite, gli ordini. Viceversa, i database operazionali sono organizzati intorno alle differenti applicazioni del dominio aziendale. La condizione di integrità e consistenza è molto importante, in quanto il DW si appoggia a più fonti di dati eterogenee: dati estratti dall'ambiente di produzione, e quindi originariamente archiviati in basi di dati aziendali, o addirittura provenienti da sistemi informativi esterni

all'azienda. Di tutti questi dati il DW si impegna a restituire una visione unificata. La costruzione di un sistema di data warehousing non comporta l'inserimento di nuove informazioni bensì la riorganizzazione di quelle esistenti, e implica pertanto l'esistenza di un sistema informativo. Infine, nel data warehouse i dati non vengono mai rimossi ma solo aggiunti, questa caratteristica consente di avere a disposizione sia dati storici che recenti. Un data warehouse può essere consultato direttamente, ma anche essere usato come sorgente per costruirne delle parziali repliche orientate verso specifiche aree dell'impresa. Tali repliche vengono dette data mart.

Nelle attuali realtà aziendali è inevitabile che vengano utilizzati sistemi diversi per produrre e memorizzare i dati necessari per lo svolgimento delle differenti attività. Si rende quindi necessario centralizzare le informazioni in un unico repository e ciò è possibile tramite un processo di data integration. Nel contesto attuale grandi quantità di dati sono memorizzate in database relazionali, tuttavia, esiste una pluralità di altre fonti di dati, più o meno strutturati (documenti di testo, fogli di calcolo, ecc.), che si è ulteriormente allargata. Per poter competere nel mercato attuale le imprese hanno la necessità di accedere alle informazioni presenti in tutte le basi dati possedute internamente, attraverso una vista unica integrata e consolidata.

Il termine Data Integration racchiude tutte quelle attività e processi che permettono di unire due o più basi di dati provenienti da fonti diverse, in un database o in una vista unica facilmente consultabile dall'utente finale, al fine di effettuare attività di analisi sui dati integrati. Nella maggior parte dei casi l'utilizzo di dati integrati permette infatti di conseguire risultati di business migliori. Gartner[2] da una definizione più articolata di Data Integration:

[2] Gartner Inc. è una società per azioni multinazionale che si occupa di consulenza strategica, ricerca e analisi nel campo della tecnologia dell'informazione. L'attività principale consiste nel supportare le decisioni di investimento dei suoi clienti in ambito ICT attraverso ricerca, consulenza, benchmarking, eventi e notizie. [Fonte Wikipedia]

"A discipline comprising the practices, architectural techniques, and tools for achieving the consistent access to, and delivery of, data across the spectrum of data subject areas and data structure types in the enterprise, in order to meet the data consumption requirements of all applications and business processes."

Secondo questo punto di vista gli applicativi di Data Integration si pongono al centro delle infrastrutture basate su dati e informazioni, garantendo la possibilità di superare i tipici problemi di condivisione dei dati e di rendere i dati fruibili per tutte le applicazioni e per i processi di business dell'azienda.

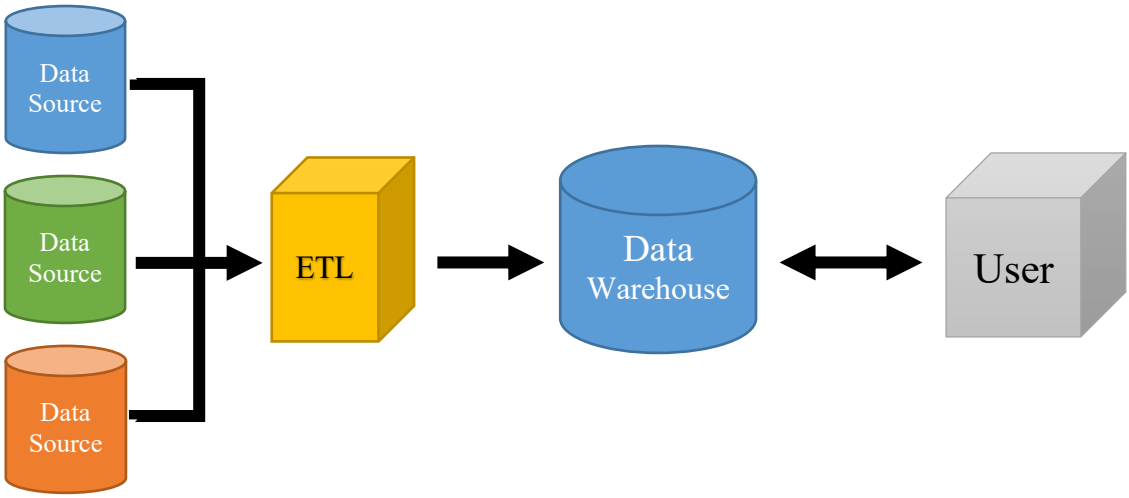

Figura 5 - Modello di creazione di un Data warehouse

2.1 I PROCESSI ETL

ETL è l'acronimo inglese di Extract, Transform and Load, le tre fasi costituenti il processo di estrazione, trasformazione e caricamento dei dati nelle diverse applicazioni destinatarie. Il ruolo degli strumenti di ETL è quello di alimentare una sorgente dati singola, dettagliata, esauriente e di alta qualità che possa a sua volta alimentare il data warehouse. Questi processi sono suddivisi in tre diverse fasi. Sta diventando sempre più comune estrarre i dati dalle sorgenti, quindi caricarli nel data warehouse di destinazione oppure trasformarli dopo il caricamento. Questo tipo di processo si chiama ELT anziché ETL.

Figura 6 - Processo ETL

2.1.1 ESTRAZIONE

L'estrazione è lo step iniziale di questo processo e rappresenta la fase di acquisizione dei dati. Le sorgenti possono essere file (CSV, JSON, XML), database transazionali o altri sistemi informatici gestionali. È possibile prendere l'intero dataset o solo una porzione. In generale, ci sono due modalità di estrazione: statica, in cui si fa una fotografia dei dati operazionali, o incrementale nella quale si preleva solo ciò che è variato dall'ultima

estrazione. L'output di questo step è l'insieme dei dati estratti, consolidati con un formato adatto al passaggio successivo.

2.1.2 TRASFORMAZIONE

La fase di trasformazione del processo ETL è la più critica. In questa fase, infatti, ai dati vengono applicate le regole aziendali necessarie a soddisfare i requisiti. La trasformazione avviene tramite l'applicazione di una serie di regole definite a livello aziendale e si rende necessaria a conformare i dati delle sorgenti alla struttura del data warehouse. La fase di trasformazione è solitamente composta da due ulteriori sottofasi: pulizia ed elaborazione. Con la fase di pulizia viene migliorata la qualità dei dati eliminando inesattezze, inconsistenze e difetti dovuti a valori errati, dati duplicati, dati mancanti. L'elaborazione ha invece lo scopo di far aderire i dati alle regole di business del sistema cui l'ETL è rivolto.

2.1.3 CARICAMENTO

L'ultima fase del processo ETL è il caricamento dei dati nella base dati di destinazione. Nella fase di caricamento esistono due modelli distinti di replicazione dei dati. Nella replicazione 'push', l'applicativo spinge i dati trasformati al database di destinazione. Nella replicazione 'pull', al contrario, l'applicazione o il database di destinazione richiedono i dati, in conformità alle esigenze specifiche del momento. I sistemi di ETL sono infatti l'infrastruttura chiave per il supporto decisionale dei sistemi di Business Intelligence. Alla fine di questa fase si ha quindi la propagazione degli aggiornamenti effettuati ed il dato viene reso disponibile al sistema target finale.

2.2 DATA QUALITY

Come descritto nei paragrafi precedenti, parte del processo di data integration riguarda l'analisi e la definizione di regole per la definizione di un database pulito senza inconsistenze ed inesattezze, sia a livello di attributo che logico.

Contrariamente a ciò che si crede i dati non sono esenti da errori e devono rispettare dei requisiti. Quando si definisce la qualità bisogna definire dei vincoli, determinare come impostare tali vincoli e determinare quale sia il livello di tolleranza degli errori. Nel corso del tempo le procedure e le tecniche si sono evolute per assicurarsi che i dati richiesti dai sistemi tradizionali possiedano un adeguato livello di qualità. Tuttavia, l'uso di dati legacy[3] ha rifocalizzato l'attenzione sulla qualità dei dati e ha messo in luce problemi come la necessità di dati soft, ovvero dati non controllati, che non si riscontrano nei sistemi tradizionali. Inoltre, i dati di oggi sono visti come una risorsa organizzativa chiave e devono essere gestiti di conseguenza. La qualità dei dati può esser dunque definita come il grado di conformità dei dati ai requisiti ad essi associati ed agli obiettivi per i quali vengono utilizzati. Il termine qualità dei dati può essere meglio descritto come "fitness for use", in modo da sottolineare come questo concetto sia relativo ai singoli contesti operativi: i dati ritenuti opportuni per uno scopo potrebbero non possedere qualità sufficienti per un altro uso. È possibile così osservare due problemi: uno legato all'accuratezza, intesa sia come accuratezza semantica che come rispondenza con il dato origine, e l'altro alla semantica.

Per quanto riguarda la prima l'idoneità all'uso implica che si deve guardare oltre le preoccupazioni tradizionali. I dati che si trovano in sistemi di tipo contabilità possono essere precisi, ma inadatti per l'uso se non sufficientemente tempestivi. Le basi di dati

3 Un sistema legacy, in informatica, è un sistema informatico, un'applicazione o un componente obsoleto, che continua ad essere usato poiché l'utente (di solito un'organizzazione) non intende o non può rimpiazzarlo. [Fonte Wikipedia]

situate in diverse divisioni di una società possono essere corrette, ma non idonee per l'uso se il desiderio è quello di combinare dati dai formati incompatibili. Un altro problema legato alla presenza di diversi utenti riguarda la semantica. I raccoglitori di dati e l'utente iniziale possono essere pienamente consapevoli delle sfumature che riguardano il significato dei vari elementi, ma che purtroppo non saranno gli stessi per tutti gli altri utenti. Così, anche se il valore può essere corretto, può essere facilmente frainteso.

In un senso molto reale i dati costituiscono la materia prima per le industrie nell'era dell'informazione. Il valore della materia prima in un'organizzazione è chiaro, almeno dal punto di vista contabile. Mentre il valore dei dati dipende quasi interamente dai suoi usi, che potrebbero anche non essere completamente noti. I nostri giorni vengono spesso definiti con il termine di era post-industriale o età dell'informazione in cui l'informazione è una forma di capitale molto importante se non addirittura la principale moneta di scambio. Pertanto, non possiamo sottovalutarne la sua qualità.

Una buona Data Governance, ovvero l'insieme di strategie, processi e regole che consentono di trattare e valorizzare i dati può comportare diversi benefici per le imprese come ad esempio:

- Un miglioramento della qualità ed una velocizzazione delle procedure di decision-making;
- Il miglioramento dell'abilità di rispondere rapidamente ai cambiamenti del mercato permettendo l'adozione di strategie di lungo termine;
- Il miglioramento della business intelligence;
- Una riduzione dei costi;
- Un aumento della customer satisfaction;
- Un miglior posizionamento sul mercato.

Il primo step per risolvere un problema è quello di essere consapevoli del problema stesso e del suo impatto, tale passo è però difficoltoso da affrontare. Per facilitare il compito di creare consapevolezza si possono mostrare le conseguenze della scarsa qualità dei dati tramite effetti più conosciuti come error rate ed errori. L'aspetto più efficace che evidenzia la necessità di Data Governance nelle aziende è rappresentato proprio dalle perdite economiche risultanti dalla scarsa Data Quality. La competitività delle imprese moderne dipende dalla loro capacità di offrire servizi personalizzati sulla base di una segmentazione della loro clientela. Il grado di personalizzazione che possono raggiungere dipende dalla qualità delle informazioni sui clienti che sono in grado di gestire, raccogliere, conservare ed estrarre dai loro database. La qualità delle informazioni è una questione fondamentale che consente all'impresa di ottenere un vantaggio competitivo basato sulla strategia customer-centric. Se le informazioni sui prodotti non incontrano le esigenze dei clienti i profitti diminuiscono.

Con il miglioramento delle informazioni sui prodotti un'organizzazione può migliorare la soddisfazione del cliente, aumentando l'efficacia e l'efficienza di un processo aziendale. Gestire la qualità delle informazioni è una attività costosa, ma la prevenzione di un errore può costare ben molto meno della sua risoluzione. Si rende così necessaria l'adozione di meccanismi atti a rilevare problemi di qualità dei dati ed a migliorare le prestazioni della qualità dei dati. Le fonti di costo dovute alla scarsa qualità sono fortemente dipendenti dal contesto e questo rende la valutazione delle perdite di scarsa qualità particolarmente difficile, come il valore dei dati stessi e la certificazione di qualità relativa ha conseguenze diverse a seconda del destinatario. Ad esempio, informazioni obsolete su uno stock possono far prendere ad un operatore decisioni di investimento sbagliate con notevoli conseguenze in termini di perdite economiche.

2.3 MISURARE LA QUALITÀ DEI DATI

Determinare il livello di qualità dei dati posseduti da un'azienda è un'operazione complessa. Per determinare la bontà dei dati è necessario definire delle metriche attraverso le quali misurare la qualità dei dati. Tuttavia, è molto difficile definire delle metriche universalmente valide in quanto la correttezza dei dati è profondamente legata ai singoli contesti operativi. La qualità del dato è un concetto multidimensionale la cui valutazione implica la definizione di metriche soggettive, adattabili ad un particolare contesto di business. È comunque possibile tentare di definire delle metriche universali indipendenti dal contesto di utilizzo dei dati. Pertanto, si possono individuare due tipologie di valutazione:

- Indipendenti dal contesto o oggettive: metriche che riflettono lo stato dei dati senza considerare come e dove vengono utilizzati;
- Dipendenti dal contesto o soggettive: misurazioni che tengono in considerazione il contesto di utilizzo, regole, caratteristiche e vincoli del business di riferimento.

Dei possibili indicatori per accertare la qualità dei dati indipendentemente dal contesto di utilizzo sono proposti da Thomas Redman[4]. Redman propone due semplici indicatori in grado di determinare il livello di correttezza di un insieme di dati:

- Correttezza a livello di attributi
- Correttezza a livello di record

Secondo Redman il livello di correttezza a livello di record è un buon indicatore di qualità della base di dati in quanto permette di identificare la percentuale di record che contengono degli errori. Tuttavia, senza tenere conto del contesto di utilizzo dei dati, tali misurazioni potrebbero risultare falsate. Altre tipologie di metriche oggettive fanno uso di

4 Thomas C. Redman, Ph.D., Data Doc and President of Navesink Consulting Group, advises organizations on their data and data quality programs.

tecniche matematico statistiche per determinare il livello di completezza e correttezza dei dati. Ad esempio, è possibile utilizzare l'analisi dell'andamento temporale dei dati per determinare gli scostamenti dal valore atteso e di identificare eventuali problematiche. La definizione di metriche in grado di considerare il contesto passa dalla definizione delle dimensioni attraverso cui valutare la qualità dei dati. Per determinare quali siano i criteri più rilevanti rispetto a cui misurare la qualità dei dati in un determinato contesto molte organizzazioni fanno compilare dei questionari agli utenti operanti nel contesto in oggetto. Le principali dimensioni da tenere in considerazione sono le seguenti:

- Accessibilità: indica la facilità con cui un utente può identificare, ottenere ed utilizzare i dati;
- Comprensibilità: determina quanto i dati sono facili da comprendere;
- Accuratezza: grado di corrispondenza fra il dato ed il reale valore della caratteristica in oggetto;
- Attendibilità: indica il grado di credibilità e affidabilità dei dati, dipende dall'attendibilità della fonte di origine;
- Completezza: è una misura di corrispondenza tra il mondo reale e il dataset specifico. Indica quanti e quali dati mancano nel dataset per offrire una rappresentazione completa al 100% del contesto reale;
- Consistenza: per ottenere una rappresentazione consistente i dati all'interno di un dataset devono essere strutturati nello stesso modo;
- Correttezza: indica il grado di esattezza e affidabilità dei dati;
- Interpretabilità: si riferisce alla disponibilità di una documentazione della base dati chiara e precisa che indichi agli utenti che tipologie di dati sono contenute nel database, come utilizzare e analizzare i dati;
- Manipolabilità: indica il grado di facilità con cui i dati possono essere elaborati per scopi differenti;

- Puntualità: indica quanto i dati sono aggiornati rispetto al contesto reale. È una misura di allineamento temporale della base dati rispetto al mondo reale e costituisce un indicatore di fondamentale importanza. Lavorare su dati obsoleti può portare a prendere decisioni critiche errate;
- Quantità: indica quanto è appropriato il volume di dati posseduti in riferimento ad una determinata attività. Lavorare con più o meno dati del necessario può rivelarsi controproducente e difficile da gestire;
- Rilevanza: capacità dell'informazione di rispondere agli obiettivi definiti; la rilevanza, o pertinenza, indica quanto i dati siano appropriati in un determinato contesto applicativo;

A partire da tali dimensioni un'organizzazione deve definire delle metriche ad hoc in grado di determinare la qualità dei dati nel proprio contesto di business.

Il processo di misurazione risulta quindi un'operazione complessa dal momento che non ci sono algoritmi univoci e precisi per il calcolo delle singole dimensioni. Esistono però algoritmi consolidati per le dimensioni di completezza, accuratezza e puntualità, o timeliness.

Completezza

Considerando un attributo in una tupla t e il suo valore v:

$$\begin{cases} se\ v = null\ \rightarrow completezza(v) = 0 \\ se\ v \neq null\ \rightarrow completezza(v) = 1 \end{cases}$$

Allora la completezza della tupla può essere calcolata come:

$$Completezza(t) = \frac{\sum_{i=1}^{N} completezza(v_i)}{N}$$

Dove N è il numero di attributi che compongono lo schema.

Accuratezza

L'accuratezza si misura considerando una sorgente di benchmark e paragonando i valori contenuti all'interno del database v_i con i valori di benchmark considerati corretti.

$$\begin{cases} se\ v_i = v_i' \rightarrow accuratezza(v_i) = 1 \\ se\ v_i \neq v_i' \rightarrow accuratezza(v_i) = 0 \end{cases}$$

L'accuratezza totale risulta:

$$Accuratezza(t) = \frac{\sum_{i=1}^{N} accuratezza(v_i)}{N}$$

Timeliness o Tempestività

Un aspetto importante dei dati è relativo al loro cambiamento e aggiornamento nel tempo. Prima di definire come si misura la tempestività, è opportuno introdurre il concetto di currency, o livello di aggiornamento, e volatilità: il livello di aggiornamento riguarda la misura in cui i dati vengono aggiornati prontamente; la volatilità definisce la frequenza di variazione dei dati nel tempo: è pari a 0 se i dati sono stabili, cioè non variano, e si esprime quantitativamente come l'intervallo di tempo in cui il dato rimane valido. Per esempio, dati stabili come le date di nascita hanno volatilità pari a 0, in quanto non subiscono nessuna variazione.

Il valore della timeliness esprime quanto siano tempestivi i dati rispetto alle esigenze temporali del contesto di utilizzo. La dimensione tempestività è motivata dal fatto che è possibile avere dati aggiornati ma inutili perché arrivati in ritardo per l'uso specifico che se ne deve fare. La tempestività è quindi ricavabile dalla seguente equazione:

$$Timeliness(v_i) = \max\left(1 - \frac{currency(v_i)}{volaitlità(v_i)}; 0\right)$$

CAPITOLO 3 – DDP, DIGITAL DEALER PLATFORM

Come esposto nei capitoli precedenti, in un contesto in forte evoluzione come quello dell'auto, stanno attualmente emergendo, per i Dealer, necessità di fornire servizi esclusivi, semplificare ed automatizzare i processi operativi ed in generale, di ottimizzare le funzioni di vendita di veicoli, accessori e servizi connessi. La risposta a queste esigenze passa attraverso la gestione, il trattamento e l'analisi di una grande quantità di dati proveniente da fonti diverse, non integrate tra loro, come i sistemi interni di proprietà (OMS, CRM, ecc.), i sistemi forniti dalle case madri relativi a tutti i brand di cui il Dealer ha il mandato o sistemi forniti da provider di servizi e altri dati esterni (es: assicurazioni, banche, siti usato, ecc.). La mancanza di integrazione e interazione tra questi sistemi crea un gap nei processi e rallenta l'esperienza del cliente.

Il quadro attuale descritto sopra che caratterizza il mondo Dealer oggi è quindi: grandi quantità di dati di forma diversa, provenienti da svariate fonti, non integrate tra loro.

Le informazioni derivanti dai dati rimangono identità singole non navigabili in maniera cross e non coinvolte in un ecosistema organico. I costi di integrazione diventano elevati, e l'aggiunta di nuove fonti genera oggi poca flessibilità e lentezza nell'adattarsi a cambiamenti dell'ambiente esterno o a nuove richieste. Spesso, il maggior rischio dell'implementazione di tali sistemi di sviluppo o integrazioni finalizzati a specifiche esigenze è quello di incorrere in investimenti non ottimizzati, che non restituiscono i benefici sperati.

Engineering Ingegneria Informatica ha implementato uno strumento a supporto del Dealer durante questa fase di trasformazione digitale. Questa società di consulenza aziendale e IT beneficia di una grande esperienza nel settore automobilistico e ricopre un'importante posizione di leader nella trasformazione digitale.

Engineering è un leader italiano nella Digital Transformation con un'offerta completa di business integration, outsourcing applicativo e infrastrutturale, soluzioni innovative e consulenza strategica.

Con circa 11.000 professionisti in 65 sedi (in Italia, Belgio, Germania, Norvegia, Repubblica di Serbia, Spagna, Svezia, Svizzera, Argentina, Brasile e Usa), il Gruppo Engineering disegna, sviluppa e gestisce soluzioni innovative per le aree di business in cui la digitalizzazione genera i maggiori cambiamenti, tra cui Digital Finance, Smart Government & E-Health, Augmented City, Digital Industry, Smart Energy & Utilities, Digital Telco & Multimedia.

La soluzione creata dalla divisione ENG4AUTO, brand con cui Engineering si rivolge agli operatori del settore automotive, a supporto dei Dealer è la Digital Dealer Platform. È possibile definire una piattaforma digitale come la raccolta di funzionalità e componenti tecnologiche attraverso le quali è possibile digitalizzare i processi o interi ecosistemi sviluppando così nuove modalità di fruizione di questi e nuovi servizi che danno valore agli utenti a partire dal dato.

Questa piattaforma digitale aiuta i Dealer nel governare le opportunità di evoluzione di un mercato in fase di cambiamento: è un sistema adibito all'integrazione di tutti i dati interni ed esterni all'ecosistema aziendale, con lo scopo di renderli fruibili per il proprio business.

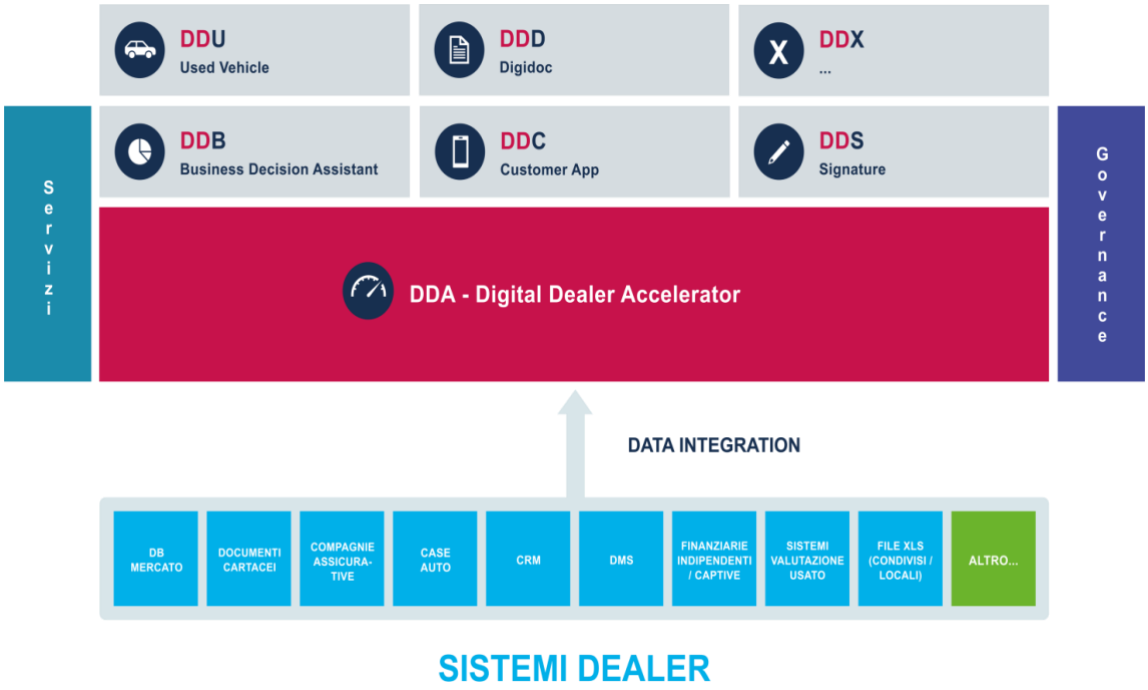

Figura 7 - Modello della struttura della Digital Dealer Platform

Come mostrato in figura 7, questa piattaforma è formata da molteplici moduli, ognuno adibito a svolgere una funzione differente in risposta alle diverse esigenze del Dealer.

I moduli da cui è composta la Digital Dealer Platform sono:

- DDA – Digital Dealer Accelerator: è il modulo abilitante agli altri della Platform. Costituisce il canale di interazione tra le diverse fonti di dati;
- DDC – Digital Dealer Customer App: è il punto di contatto tra il Dealer ed i clienti e permette una comunicazione diretta tra le parti attraverso la generazione di notifiche segmentate, aumentando la riconoscibilità della concessionaria;
- DDB – Digital Dealer Business Decision Assistant: è una piattaforma integrata che, grazie ad una interfaccia grafica di facile lettura permette al Dealer di gestire e monitorare l'intero business. Questo modulo è strutturato attraverso l'unione di analisi predittive, processi di clustering e funzioni analitiche e facilita, ad esempio, la gestione dello stock, il monitoraggio delle catene di vendita e la valutazione

delle performance dei venditori. Nell'ottica di una gestione integrata dell'attività, ogni dato contiene informazioni preziose per il business che, se riorganizzati nei giusti KPI, diventano elemento essenziale per monitorare le performance del proprio business e per supportare la simulazione e la successiva assunzione di decisioni operative "complesse. Una volta riaggregati i dati inerenti alle attività aziendali, gli analytics offrono così una reportistica operativa e una dashboard per il management;

- DDS – Digital Dealer Signature: processo paperless che utilizza la firma elettronica avanzata per lo snellimento dei processi e archiviazione elettronica;
- DDD – Digital Dealer Digidoc: servizi specifici per lo sviluppo digitale e la gestione e archiviazione della documentazione;
- DDU – Digital Dealer Used Vehicle: portale di gestione dell'usato dalle fasi della perizia fino alla vendita del veicolo.

3.1 DDA – DIGITAL DEALER ACCELERATOR

Una soluzione strutturata come la DDP richiede la concentrazione di molte componenti software, ognuna delle quali ha un ambito tecnologico differente; il cuore di questa soluzione (DDA – Digital Dealer Accelerator) si basa sulla raccolta di dati da fonti eterogenee e sulla integrazione di questi dati in un insieme di informazioni coerenti.

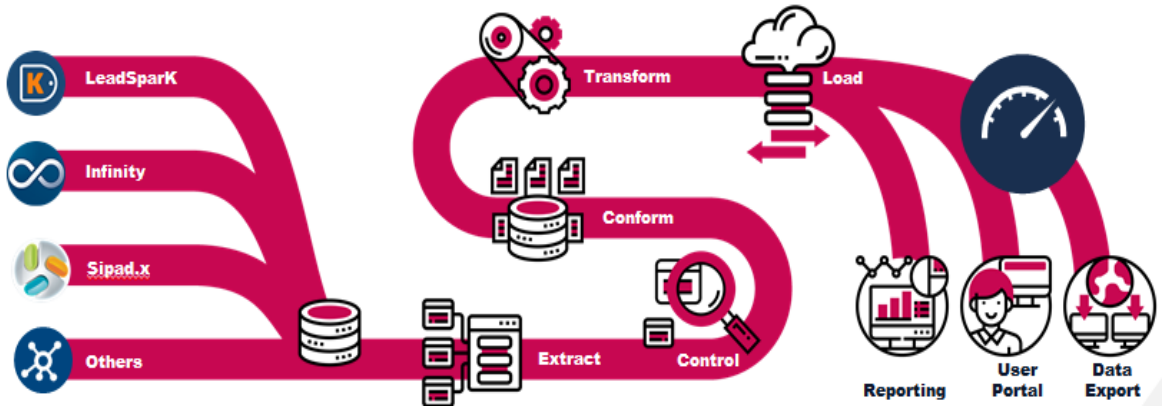

Figura 8 – Modello del processo ETL all'interno della DDP

Tramite questo modulo centrale, le informazioni che risiedono nei diversi sistemi utilizzati (es. DMS, CRM, sistemi forniti dalle case auto, dati provenienti dalle società finanziarie o dalle assicurazioni ecc.), vengono collezionate per poter essere aggregate in un unico punto. I dati vengono inizialmente catturati dai vari data sources e, una volta raccolti, vengono processati per escludere duplicati ed informazioni estranee. I data sources sono i connettori in ingresso al Data Hub centrale. I dati vengono trasferiti tramite protocolli di trasferimento dati standard: S/FTP e HTTP/S. Le sorgenti dati possono essere:

- Fonti dati dirette (DB);
- Estrazione di dati da sistemi esterni (File CSV, XML, XSL);
- Servizi messi a disposizione per recuperare i dati on demand (WS, API).

I dati vengono così resi accessibili ai sistemi del cliente o agli add-on del DDA, dopo esser stati trasformati e caricati tramite un processo di data integration. Sono previsti

anche servizi di export dei dati chiamando sistemi esterni o esponendo API verso i sistemi esterni.

La fase successiva è la fase di data aggregation, in cui vengono analizzati i dati ed aggregati secondo le regole definite con il cliente.

In caso di dato proveniente da più fonti vengono precedentemente definite le priorità e vengono caricati i dati nelle entità proprie del Data Hub elencate sopra. I file provenienti di sistemi esterni vengono copiati nel repository centrale, ovvero un archivio digitale che raccoglie e visualizza i set di dati e i relativi metadati. Successivamente vengono effettuati controlli formali sui file e sul loro contenuto: viene infatti analizzata la nomenclatura, l'integrità ed il tracciato.

La parte di import dati ed integrazione è costruita attraverso l'utilizzo dello strumento Talend, uno dei prodotti leader di mercato per la data integration; le operazioni di caricamento ed integrazione vengono eseguiti da un ambiente di runtime Java 1.8. La tecnologia RDBMs è MariaDB, un database open source nato a partire dalla code-base di MySQL e sviluppato dai programmatori originari di MySQL.

3.2 I PRINCIPALI SISTEMI OPERAZIONALI DI UN DEALER

I sistemi operazionali svolgono la funzione principale di gestione di tutti i flussi informativi che riguardano le operazioni quotidiane della gestione aziendale.

I cambiamenti descritti precedentemente si stanno verificando in un contesto dove la mentalità di gestione dei Dealers deve saper evolvere da un approccio di business puramente familiare ad un approccio strettamente manageriale. Questo approccio family business ha portato ad una mancanza di una programmazione a lungo termine ed il

sistema adottato riflette, specialmente nei grandi gruppi Dealer, questa improvvisazione manageriale.

Il cuore del sistema informativo per ogni concessionario è il Dealer Management System, anche conosciuto come DMS. Questo sistema di gestione è proprio dei concessionari d'auto ed è una categoria di software finalizzato alla gestione di tutte le attività tipiche, come la gestione dei punti vendita e dei magazzini, dei diversi marchi distribuiti con le loro interfacce e specificità e la gestione delle vendite, sia di macchine nuove che usate, con gli eventuali servizi di finanziamento e servizi post-vendita. Gli attuali DMS si integrano quindi con i cataloghi e i sistemi di contratto dei vari OEM per la gestione delle vendite delle auto nuove.

In Italia sono presenti sei principali fornitori di DMS, esposti nella tabella seguente.

DMS Provider	DMS
CDKGlobal	CDK DMS
Pentana Solutions (Esseitalia)	Sipad.X
Porsche Informatik	CROSS2
Incadea	Incadea DMS
Visual Software	Infinity
Global Automotive	DataCar DMS, Concerto

Tabella 1 - Provider DMS

Ovviamente, il DMS non è perfettamente sviluppato in ogni parte e soddisfa bisogni generali e non specifici. Per questa ragione, un dealer può decidere di utilizzare un DMS per alcune attività e di implementare un altro sistema per svolgerne altre.

I costi di transazione e di switch per implementare un nuovo DMS sono molto alti e possono creare situazioni difficili da gestire. Le differenze tra i DMS possono essere alla base dei difficili problemi di integrazione. Inoltre, i costruttori di auto spesso obbligano i concessionari ad implementare ed adottare le loro applicazioni per specifiche attività come i contratti di vendita, servizi finanziari e garanzie.

I DMS, inoltre, non riescono a raggiungere lo scopo di una buona lead management con la stessa intensità e capacità che può fornire un'applicazione CRM specializzata e lo stesso vale per i servizi finanziari. Il sistema informativo risultante che si ottiene può esser visto quindi come un enorme silos pieno di dati ed informazioni ma di difficile esplorazione per la creazione di valore aggiunto. La struttura descritta mostra già un importante grado di complessità ma, quando un dealer acquisisce un altro concessionario, il problema cresce sensibilmente: ogni concessionario ha una sua soluzione centrale, fornito da aziende diverse ed ogni fornitore utilizza propri codici. Ciò comporta una nuova codifica dei dati e degli attributi in un nuovo repository.

3.3 DAL SISTEMA OPERAZIONALE AL DATA HUB: DATA INTEGRATION

Nel paragrafo 3.1 è stato descritto il processo generale di integrazione dei dati dai diversi data source al data hub centrale. Nel seguito verrà descritta la modalità con cui questi dati vengono integrati.

I Data Source possono fornire dati secondo due modalità:

1. Web Service, con cui il Data Source fornisce i dati in maniera Push (il Data Source invoca il Sistema destinatario e gli fornisce i dati) oppure in maniera Pull (il Data Source è in ascolto e si aspetta di essere interrogato dal Sistema che abbisogna dei dati); la fornitura di dati attraverso Web Service è sempre "puntuale": il Data Source invoca un altro Sistema e gli invia un "pacchetto di dati" (es. l'ultima transazione registrata dal Sistema Operazionale);
2. Flat file (es. CSV, TXT, JSON, etc.). con cui il Data Source fornisce i dati attraverso un Flat-file a tracciato record fisso depositato (es. via SFTP) in una

cartella condivisa tra Data Source e Sistema di destinazione; la fornitura in modalità Flat file può avvenire in modalità "Delta" o in modalità "Snapshot":

a. Flatfile.Delta: il Flat-file ospita le variazioni intervenute nei dati dell'Entità veicolata nel Data Source in un intervallo di tempo (es. il Flat file ospita i Contratti di Finanziamento inseriti nella giornata del gg/mm/aaaa);
b. Flatfile.Snapshot: il Flat-file ospita una fotografia dei dati dell'Entità veicolata nel Data Source, fotografia scattata nell'istante in cui il Data Source espone i dati.

La modalità di integrazione dati utilizzata ed implementata nella soluzione Digital Dealer Platform è il trasferimento dati tramite Flat file a tracciato fisso in formato CSV (o analogo) con dati codificati in ASCII UTF8 (utf8mb4_bin). Si ipotizza l'impiego di un diverso Flatfile per ciascun Data Source che debba alimentare il Data Hub.

Si richiede che la componente di Data Collection abbia la capacità di catturare i Flat file che ospitano i Data Source, dall'area nella quale li ha depositati il Sistema Operazionale, attraverso un meccanismo di trasporto sicuro e protetto (es. SFTP) governato dalla componente di Data Collection stessa.

La componente di Data Collection si pone quindi in maniera "attiva" (con modalità "polling") rispetto al Sistema Operazionale e possa catturare in autonomia il Flat file (che si assume il Sistema Operazionale abbia prodotto nel rispetto della frequenza stabilita).

La componente di Data Integration accetta tutto ciò che proviene dai Data Source e non produce scarti perché durante la vita della soluzione adottata dal cliente non sono previsti servizi o persone in grado di gestirli.

Viene definito come scarto un record proveniente da un Data Source al cui interno sono presente dei campi il cui valore non è conforme a quello atteso dal Data Hub. Esempi di scarti sono i record in cui un campo obbligatorio è privo di valore, quelli in cui un campo

ha un valore fuori dal dominio dei valori ammissibili o quelli in cui due campi tra loro relazionati non sono coerenti (es. valore del campo "Località" non coerente con il valore del campo "Provincia"). Vengono quindi gestiti solo quei record non corretti dal punto di vista del formato. Nessuna attenzione viene posta sulla relazione logica e sintattica dei dati, generando così possibili ridondanze e rumori, considerando anche l'eterogeneità delle fonti da cui questi dati vengono integrati.

CAPITOLO 4 – ANALISI E DEFINIZIONE DI INDICATORI PER LA MISURA DELLA QUALITÀ DEL DATO

4.1 INTRODUZIONE

Lo scopo di questo capitolo è di presentare il lavoro di analisi svolto durante il periodo di tirocinio curriculare ed il successivo periodo di inserimento in azienda. Per motivi di privacy non verrà nominato il nome del dealer protagonista dell'oggetto di studio, e verrà semplicemente chiamato Dealer.

Per venire incontro alle esigenze del Dealer, Engineering Ingegneria Informatica ha sviluppato per il cliente la piattaforma descritta nel capitolo precedente, la Digital Dealer Platform.

A causa della numerosità di fonti di dati, si è reso necessario realizzare un repository centrale in grado di ricevere ed effettuare aggregazioni dei dati da più fonti. La soluzione implementata per aggregare, archiviare e utilizzare i dati dei differenti sistemi operativi in uso è chiamato "Dealer Data Hub". Il Data Hub offre omogeneità e persistenza storica ai dati e li rende pienamente fruibili sia per la comunicazione verso il Driver che per l'analisi svolte dal modulo di Business Intelligence.

Il sistema informativo per la gestione del concessionario adottato dal Dealer è Sipad.X fornito da Esseitalia (ora acquisito da Pentana Solutions). Le attività che vengono svolte facendo uso di questo DMS (Dealer Management System) sono:

- Vendite di veicoli nuovi;
- Vendite di veicoli usati;
- Servizio di riparazione e / o manutenzione;

- Vendite di pezzi di ricambio;
- Gestione stock di veicoli nuovi;
- Gestione stock di veicoli usati.

Oltre a questo DMS, ci sono molte altre soluzioni adottate dal Dealer per la conduzione delle sue attività:

- Sistemi Operazionali impiegati per la conduzione delle attività quotidiane nella gestione delle vendite e i servizi della carrozzeria;
- Portali dei Brand impiegati per coadiuvare l'attività di Sales, After-Sales ed eventuale Rent;
- Portali di Provider impiegati per la gestione dei finanziamenti e delle assicurazioni (es. Findomestic)

Prima di approfondire l'argomento di data quality inerente il caso di studio e di poter così definire degli indicatori di qualità del dato, è necessario descrivere la struttura logica del database utilizzato per le analisi in modo da poter identificare le principali entità di analisi.

Il DB è formato da quattro principali schemi:

- Data Hub centrale
- Data Mart
- Tabelle di staging
- Data Config

Sarà oggetto dell'analisi solo il Data Hub, poiché gli altri schemi sono funzionali al corretto utilizzo di quest'ultimo, che rappresenta il repository centrale.

Le principali entità definite nel Data Model che costituisce il Data Hub sono:

- Soggetto
- Veicolo
- Offerta vendita nuovo e usato
- Contratto vendita nuovo e usato
- Stock veicoli
- Fatture attive e passive
- Ordini vettura
- Offerta aftersales
- Ordine di lavoro
- Attività officina
- Magazzino ricambi
- Offerta Sales Garanzie e Assicurazioni

Le entità sono tra loro interconnesse per consentire di descrivere al completo una specifica informazione tematica. Nella figura successiva viene proposta una visualizzazione delle connessioni presenti tra le varie entità.

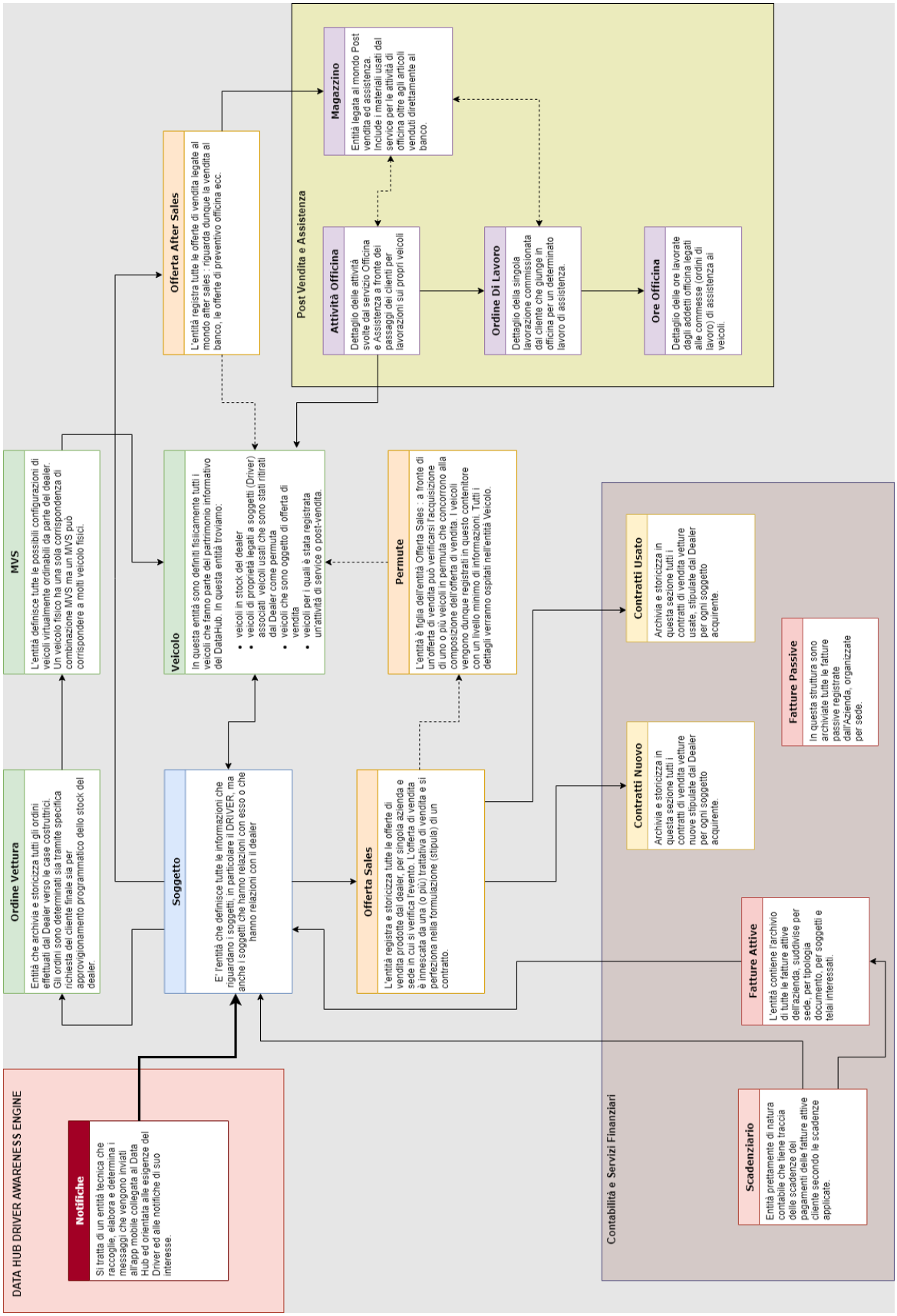

Figura 9 - Modello delle connessioni tra entità

Al centro del patrimonio informativo ci sono le entità relative ai soggetti ed ai veicoli. Attorno a queste macro entità ruotano tutte le informazioni collegate ed utilizzate per i principali scopi previsti dalla piattaforma: lato Driver, le informazioni sono messe a disposizione per l'app mobile; lato Dealer, queste informazioni confluiscono ed arricchiscono la piattaforma Analytics di Business Intelligence prodotta e sono utilizzati in altri servizi specifici sviluppati per il Dealer.

Si è scelto di concentrare le analisi sulle entità centrali del veicolo e del soggetto.
Nell'entità veicolo sono definiti fisicamente tutti i veicoli che fanno parte del patrimonio informativo del Data Hub. In questa entità troviamo:

- Veicoli in stock del Dealer;
- Veicoli di proprietà legati a soggetti (Driver);
- Veicoli usati ritirati dal Dealer come permuta;
- Veicoli oggetto di offerta o vendita;
- Veicoli per i quali è stata registrata un'attività di service o post-vendita.

Le informazioni relative a questa entità sono salvate, all'interno del Data Hub, nella tabella anag_base_daw, dove sono stati raccolti tutti i singoli codici provenienti da Sipad.X e suddivisi in base alla marca, al modello e alle configurazioni sia degli interni che della carrozzeria che il veicolo supporta.
Gli attributi che compongono la tabella anag_base_daw sono i seguenti:

- ID (int): numero intero incrementale;
- COD_DAW (varchar): codice creato dall'aggregazione del campo ID e del campo COD_TIPO_ANAG_BASE;
- COD_TIPO_ANAG_BASE (varchar): codice identificativo della tipologia a cui si riferisce il dato, per esempio se il campo riporta la scritta COLEST significa che il

codice si riferisce al colore degli esterni del veicolo, oppure se il campo riporta la scritta MARCA significa che il codice identifica una marca del veicolo presente nel sistema informativo del Dealer;

- COD_DS (varchar): codice utilizzato per identificare il record nel sistema informativo sorgente del Dealer;
- COD_SOURCE (varchar): descrizione della fonte da cui è stato estratto il dato (ad esempio, nel caso preso in esame, Sipad.X);
- DESCRIPTION (varchar): descrizione del codice;
- CREATION_DATE e LAST_UPDATE_DATE (timestamp): data di creazione e ultima modifica del dato.

L'entità soggetto definisce tutte le informazioni che riguardano i soggetti, in particolare i Driver, ma anche quei soggetti che hanno relazioni con il Dealer. Le informazioni relative a questa entità sono salvate, all'interno del Data Hub, nella tabella soggetto. La struttura della tabella è riportata in figura 10 e 11, dove vengono mostrate le colonne in cui è presente un vincolo di chiave e non tutta la totalità, poiché questa tabella contiene 318 colonne.

#	Nome	Tipo di dati	Lunghezza/set
1	**RECORD_ID**	**BIGINT**	**20**
2	SOG_ID	TEXT	
3	**ACC_SPE_IND**	**TINYINT**	**1**
4	**ACC_SPE_TS**	**TIMESTAMP**	
5	TENANT_ID	SMALLINT	6
6	GRUPPO_ID	SMALLINT	6
7	AZIENDA_ID	SMALLINT	6
8	SOG_SISOPE_COD	TEXT	
9	COGNOME	TEXT	
10	NOME	TEXT	
11	NOME_SECONDARIO	TEXT	
12	GENERE	TEXT	
13	FORMULA_SALUTO	TEXT	
14	ORGA_DEN	TEXT	
15	ORGA_DEN_EXT	TEXT	
16	ORGA_FORMA_LEG	TEXT	
17	ORGA_CLASSE_DIP	VARCHAR	8
18	ORGA_CLASSE_FAT	VARCHAR	8
19	ORGA_CLASSE_DULTAG	DATE	

Figura 10 - Subset colonne della tabella data_hub.soggetto

#	Nome	Tipo di dati	Lunghezza/set
21	ORGA_SETT_MERC_LAST_DATE_UPDATE	VARCHAR	8
22	ORGA_OPERATIVA	VARCHAR	8
23	ORGA_OPERATIVA_DATA	DATE	
24	NATURA_SOGG_COD	VARCHAR	8
25	NATURA_SOGG_DES	TEXT	
26	PERS_CODFIS	VARCHAR	32
27	LINGUA_COD	VARCHAR	8
28	LINGUA_DES	VARCHAR	8
29	PERS_CITTADINANZA_COD	VARCHAR	30
30	PERS_TIPO_SOGGETTO_COD	VARCHAR	8
31	PERS_TIPO_SOGGETTO_DES	TEXT	
32	ORGA_ID_FISCALE_COD	VARCHAR	32
33	ORGA_ID_FISCALE_DES	VARCHAR	32
34	ALTERNATIVE_KEY_SOGG	TEXT	
35	PERS_DOCUM_IDENT_TIPO	TEXT	
36	PERS_DOCUM_IDENT_ENTE	TEXT	
37	PERS_DOCUM_IDENT_DATARIL	DATE	
38	PERS_DOCUM_IDENT_DATASCAD	DATE	

Figura 11 - Subset colonne della tabella data_hub.soggetto

4.2 DEFINIZIONE DEL PROCESSO DI ANALISI

L'analisi della qualità dei dati è stata condotta tramite un processo di indagine definito a monte dell'analisi vera e propria. Il processo è composto da cinque step nei quali: inizialmente si cerca di identificare gli attributi più significativi e ne viene analizzata

l'accuratezza. Viene poi effettuata un'analisi sui record poco accurati per vedere la loro incidenza e cardinalità all'interno dei contratti nuovi e usati e il peso economico all'interno delle fatture, per quanto riguarda l'entità veicolo e viene fatta un'analisi di ricerca di pluralità tra l'entità soggetto. Successivamente a queste fasi si raggiunge alla definizione di metriche di misura della qualità del dato, all'interno di un database alimentato da fonti eterogenee utilizzate dal Dealer.

Il processo di analisi è così definito:

1. Identificazione degli attributi più significativi ed analisi dell'accuratezza del campo relativo alla descrizione;
2. Conteggio dei record con pari descrizione (entità veicolo), conteggio dei record con pari codice fiscale o partita iva (entità soggetto);
3. Analisi dell'utilizzo del codice all'interno dei processi di contratto vendita nuovo, usato e fatture attive (entità veicolo)
4. Definizione di metriche di misura della qualità del dato nel contesto sopra descritto;
5. Analisi del valore che l'informazione assume all'interno del database: le informazioni legate agli attributi selezionati hanno senso all'interno del contesto analizzato? Un'eventuale azione di ripulitura porterebbe ad un effetto potenzialmente negativo? Esiste un metodo per preservare l'informazione aggregata?

4.3 ANALISI DELL'ENTITÀ VEICOLO ALL'INTERNO DEL DATABASE

Seguendo i passi del processo di analisi definito nel paragrafo precedente, il primo step riguarda l'identificazione degli attributi più significativi e l'analisi dell'accuratezza del

campo relativo alla descrizione. Eseguendo la query sotto riportata, sono state identificate tutte le tipologie di codice utilizzate nel sistema del Dealer.

```
SELECT DISTINCT (COD_TIPO_ANAG_BASE)
FROM anag_base_daw
```

COD_TIPO_ANAG_BASE	DESCRIZIONE ATTRIBUTO
ALIM	Tipologia di alimentazione del motore
ANTINQ	Classificazione Euro dei veicoli rispetto alla normativa anti-inquinamento
AZIENDA	Azienda parte del gruppo del Dealer
CATEGOM	Codici utilizzati nell'ambito officina
CLCONT	Classe contabile
COLEST	Colore carrozzerie esterna
COLINT	Colore interni del veicolo
DESTCOM	Destinazione commerciale
FSCONTO	Fasce di sconto
GRPSTCK	Gruppo stock
INT	Descrizione degli interni del veicolo
LINEA	Linea del veicolo
MARCA	Marca del veicolo
MARCHIO	Marca del veicolo
MODEL	Modello del veicolo
PAGAM	Tipologia di pagamento
SEDE	Sede in cui è stata effettuate un'operazione relativa al veicolo
SEGM	Segmento
STAVEND	Stato della vendita
TIPART	Tipologia articolo
TIPVEND	Tipologia vendita
VEND	Venditore
VERS	Versione veicolo

Tabella 2 - Risultato query codici distinti anagrafica

I codici riportati in tabella 2 hanno tutti come sorgente il sistema DMS Sipad.X.

I record inerenti al veicolo sono: ALIM, ANTINIQ, COLEST, COLINT, INT, LINEA, MARCA, MARCHIO, MODEL.

Nella tabella anag_base_daw non esiste nessun attributo che tenga traccia o che indichi il processo in cui il relativo codice viene utilizzato.

Si è deciso di procedere con un'analisi più approfondita del campo "MARCA". Questo campo risulta di particolare interesse poiché viene utilizzato nella maggior parte delle entità definite nel data model e in molti KPI definiti nella Business Intelligence. Dalla tabella anag_base_daw sono stati estratti tutti i record in cui il cod_tipo_anag_base risulta uguale a "MARCA" ed effettuata un'operazione di conteggio dei record aventi il campo DESCRIPTION uguale.

COUNT(COD_DS)	DESCRIPTION
4	BMW
4	FIAT
4	JEEP
3	ALFA ROMEO
3	DAEWOO
3	JAGUAR
3	LANCIA
3	LAND ROVER
3	MERCEDES
3	MG
3	MINI
3	NISSAN
3	PEUGEOT
3	RENAULT
3	ROVER
3	SMART
3	VOLKSWAGEN
3	VOLKSWAGEN VIC
3	VOLVO
2	ABARTH
2	ASTON MARTIN
2	AUDI
2	BENTLEY
2	CHRYSLER
2	CITROEN
2	DACIA
2	DAIHATSU
2	DODGE
2	DS
2	FORD
2	HONDA

COUNT(COD_DS)	DESCRIPTION
2	HUMMER
2	HYUNDAI
2	INFINITI
2	ISUZU
2	IVECO
2	KIA
2	LADA
2	LEXUS
2	LOTUS
2	MAN
2	MASERATI
2	MAZDA
2	MITSUBISHI
2	OPEL
2	PORSCHE
2	SAAB
2	SEAT
2	SKODA
2	SUBARU
2	SUZUKI
2	TATA
2	TOYOTA
1	AC
1	AIXAM
1	AIXAM MOTO
1	ALPINA-BMW
1	ARCTIC CAT
1	ARO
1	AUSTIN HEALEY
1	AUTOBIANCHI
1	BARTOLETTI

COUNT(COD_DS)	DESCRIPTION
1	BENELLI MOTO
1	BMW MOTO
1	BNW
1	CADILLAC
1	CARMOSINO
1	CHATENET MOTO
1	CHEVROLET
1	CHEVROLET (K)
1	COMAI
1	CORVETTE
1	DAF
1	DR
1	DR MOTOR
1	FERRARI
1	GAC GONOW
1	GA200 2.0 GARAGE ITALIA
1	GIOTTI VICTORIA
1	GONOW
1	GREAT WALL
1	GREAT WALL_MOTOR
1	GRECAV
1	HARLEY DAVIDSON MOTO
1	HONDA MOTO
1	IMP. S.A.L.E.
1	IMPORTAZIONE
1	ITALCAR
1	JDM
1	KAWASAKI MOTO
1	LAMBORGHINI
1	LIGIER
1	LIGIER MOTO
1	MAHINDRA
1	MARCHIO GENERICO
1	MERCEDES V.C.
1	MERCEDES V.I.
1	MICROCAR
1	MICROCAR MOTO
1	MINERVA
1	MITSUBISHI AUTO
1	MITSUBISHI FUSO/CANTER
1	MUSTANG
1	MV AGUSTA
1	MV AGUSTA MOTO
1	PETTENELLA
1	PIAGGIO
1	PIAGGIO MOTO
1	PIAGGIO MOTOFURGONI
1	PUMA ITALIA
1	RAYTON FISSORE
1	RENAULT TRUCKS
1	SCANIA
1	SCANIA CV 480
1	SCANIA CV R 420
1	SCANIA R500
1	SUZUKI MOTO
1	TRIUMPH
1	TRIUMPH MOTO
1	TRUCK - V.I.
1	VAN - V.C.
1	VEM
1	VIBERTI
1	WOLKSVAGEN
1	YAMAHA
1	YAMAHA MOTO
1	ZK

Tabella 3 - Risultato query codici marca multipli

Dai risultati riportati in tabella si nota come più codici distinti siano associati alla medesima marca. È possibile osservare, inoltre, come alcuni record all'interno del

campo descrizione non corrispondono a marche reali, ma sono dei modelli o sono dati da refusi: si veda ad esempio il record "WOLKSVAGEN" o "BNW", evidenti errori di battitura. Dalla tabella sopra risultano esserci 203 record distinti relativi alle marche dei veicoli, di cui 54 sono marche a cui sono associati codici multipli e 24 sono state definite come non marche o come errori di battitura. Quest'ultimi record sono stati trovati effettuando un semplice confronto e cercando un riscontro del valore del campo con valori trovati sul web. Non è stato possibile effettuare un confronto diretto poiché non si disponeva di una base dati contenente tutte le marche di veicoli esistenti ad oggi e quindi è stata effettuata una ricerca manuale per verificare l'accuratezza dell'attributo.

Sia per le marche che presentano codici multipli, che per le marche che sono state definite come errori di battitura o che non risultano essere delle marche, per vedere quali di questi codici fossero i più utilizzati all'interno delle basi dati relative ai contratti di vendita nuovo e usato e alle fatture, sono state effettuate delle interrogazioni sulle tre tabelle di interesse: contratto_ven_nuovo, contratto_ven_usato, fatture_att_testata. Non verrà presentata la struttura delle tre tabelle poiché composte da un numero molto elevato di attributi e di non interesse per le analisi effettuate.

I record che risultano essere errori di battitura o che non risultano essere delle marche sono:

COD_DS	DESCRIPTION
FGGRFO	BARTOLETTI
SQWODK	BNW
CHA-M	CHATENET MOTO
CHC	CHEVROLET (K)
DRM	DR MOTOR
GAC	GAC GONOW GA200 2.0
BOLHPQ	GARAGE ITALIA
2715	GREAT WALL MOTOR
IMP	IMPORTAZIONE
LIG-M	LIGIER MOTO
99	MARCHIO GENERICO
VQHTC	MERCEDES V.C.

COD_DS	DESCRIPTION
36	MERCEDES V.I.
766	MICROCAR
MIC-M	MICROCAR MOTO
78	MITSUBISHI FUSO/CANTER
100	PETTENELLA
PIA-M	PIAGGIO MOTO
PMF-M	PIAGGIO MOTOFURGONI
LWDGDR	SCANIA CV 480
CQJXFW	SCANIA CV R 420
MFOSRP	SCANIA R500
76	TRUCK - V.I.
77	VAN - V.C.
1	WOLKSVAGEN

Tabella 4 - Risultato query marche errate

Su questi record sono state effettuate ulteriori analisi andando a conteggiare le righe, nelle tre tabelle definite sopra, in cui questi codici sono presenti e andando a calcolare la percentuale sul totale, in modo da evidenziarne l'incidenza.

TABELLA	TOTALE contratti	Numero di contratti in cui sono presenti marche errate	%
CONTRATTO_VEN_NUOVO	224059	2305	1,0%
CONTRATTO_VEN_USATO	359979	279	0,1%

Tabella 5 - Conteggio del numero di contratti in cui sono presenti marche errate

TABELLA	VALORE ECONOMICO TOTALE fatture	VALORE ECONOMICO delle marche errate	%
FATTURE_ATT_TESTATA	2.840.730.588,40 €	92.390.963,13 €	3%

Tabella 6 - Valore economico delle fatture in cui sono presenti marche errate

Dalle tabelle si può vedere come questi codici abbiano un'incidenza piuttosto ridotta sul totale dei contratti e sulle fatture presenti nel database. Dopo aver calcolato l'incidenza sul totale, sono state effettuate ulteriori interrogazioni per poter avere un'analisi di dettaglio e per vederne l'incidenza non più sul totale, ma l'incidenza dei singoli codici sul sottoinsieme definito.

```
SELECT COUNT(MARCA_SISOPE_COD),MARCA_SISOPE_COD
FROM CONTRATTO_VEN_NUOVO
WHERE ACC_SPE_IND=1 AND MARCA_SISOPE_COD IN ("FGGRFO" , "CHA-M",
"CHC", "DRM", "GAC", "BOLHPQ", "2715", "IMP", "LIG-M", "99", "VQHTC",
"36", "766", "MIC-M", "78", "100", "PIA-M", "PMF-M", "LWDGDR",
"CQJXFW", "MFOSRP", "76", "77", "1")
GROUP BY MARCA_SISOPE_COD
```

COUNT(MARCA_SISOPE_COD)	MARCA_SISOPE_COD	MARCA_SISOPE_DES
714	76	Truck - V.I.
1378	77	Van - V.C.
209	78	Mitsubishi Fuso
4	99	Marchio generico

Tabella 7 - Risultato query dell'analisi marche errate in relazione al contratto di vendita nuovo

```
SELECT COUNT(MARCA_SISOPE_COD),MARCA_SISOPE_COD
FROM CONTRATTO_VEN_USATO
WHERE ACC_SPE_IND=1 AND MARCA_SISOPE_COD IN ("FGGRFO" , "CHA-M",
"CHC", "DRM", "GAC", "BOLHPQ", "2715", "IMP", "LIG-M", "99", "VQHTC",
"36", "766", "MIC-M", "78", "100", "PIA-M", "PMF-M", "LWDGDR",
"CQJXFW", "MFOSRP", "76", "77", "1")
GROUP BY MARCA_SISOPE_COD
```

COUNT(MARCA_SISOPE_COD)	MARCA_SISOPE_COD	MARCA_SISOPE_DES
1	100	PETTENELLA
6	2715	GREAT WALL MOTOR
46	36	MERCEDES V.I.
4	76	CADILLAC
62	766	MICROCAR
1	BOLHPQ	GARAGE ITALIA
3	CHA-M	CHATENET MOTO
105	CHC	CHEVROLET
1	CQJXFW	SCANIA CV R 420
3	DRM	DR MOTOR
1	FGGRFO	BARTOLETTI
4	GAC	Gac Gonow GA200 2.0
6	LIG-M	LIGIER MOTO
1	LWDGDR	SCANIA CV 480
1	MFOSRP	SCANIA R500
1	MIC-M	MICROCAR MOTO
6	PIA-M	PIAGGIO MOTO
1	PMF-M	PIAGGIO MOTOFURGONI
25	VQHTC	Van - V.C.

Tabella 8 - Risultato query dell'analisi marche errate in relazione al contratto di vendita usato

```
SELECT SUM(IMP_TOT_FATTURA), COUNT(MARCA_SISOPE_COD),MARCA_SISOPE_COD
MARCA_SISOPE_DES, TELAIO_COD
FROM FATTURE_ATT_TESTATA
WHERE ACC_SPE_IND=1 AND MARCA_SISOPE_COD IN ("FGGRFO" , "CHA-M",
"CHC", "DRM", "GAC", "BOLHPQ", "2715", "IMP", "LIG-M", "99", "VQHTC",
"36", "766", "MIC-M", "78", "100", "PIA-M", "PMF-M", "LWDGDR",
"CQJXFW", "MFOSRP", "76", "77", "1")
GROUP BY MARCA_SISOPE_COD
```

SUM(IMP_TOT_FATTURA)	COUNT(MARCA_SISOPE_COD)	MARCA_SISOPE_COD	MARCA_SISOPE_DES
2.950,00 €	1	PIA-M	PIAGGIO MOTO
3.500,00 €	1	CQJXFW	SCANIA CV R 420
3.677,00 €	1	GAC	Gac Gonow GA200 2.0
5.000,00 €	1	CHA-M	CHATENET MOTO
6.300,00 €	1	MIC-M	MICROCAR MOTO
13.200,00 €	1	2715	GREAT WALL MOTOR
19.520,00 €	1	MFOSRP	SCANIA R500
150,00 €	1	PMF-M	PIAGGIO MOTOFURGONI
175.000,00 €	1	100	PETTENELLA
6.900,00 €	2	VQHTC	Van - V.C.
4.180,00 €	3	DRM	DR MOTOR
78.026,26 €	8	766	MICROCAR
818.072,00 €	24	36	MERCEDES V.I.
135.348,30 €	30	CHC	CHEVROLET
6.511.121,78 €	169	78	Mitsubishi Fuso/Canter
55.721.085,55 €	563	76	Truck - V.I.
28.886.932,24 €	772	77	Van - V.C.

Tabella 9 - Risultato query dell'analisi marche errate in relazione alle fatture

Dopo aver effettuato un'analisi su quei codici a cui è associato un campo descrizione errato, è stata spostata l'attenzione sui codici che risultano essere multipli rispetto all'unico campo descrizione della marca. Da una ricerca tra i flussi informativi raccolti dall'azienda nelle fasi di progettazione del data hub, è emerso che, nel suo sistema gestionale, il Dealer presenta più codici relativi alla medesima marca poiché questi codici vengono utilizzati in processi differenti, ed in particolare viene utilizzato un codice nelle operazioni di officina, un codice per i contratti di vendita di auto nuove ed un codice per i contratti relativi alle auto usate.

Dalle tabelle relativi ai contratti e alle fatture sono state estratte le marche più utilizzate e, come per le marche errate, è stato effettuato un conteggio ed è stato calcolato il valore economico delle fatture in cui questi codici sono stati usati.

COD_DS	# in contratto_ven_nuovo	# in contratto_ven_usato	# fatture_att_testata	Valore economico
UMVDIB 30	24607		3672	154.634.552,08 €
85		20502	3053	86.262.146,47 €
BMW		6297	1171	25.519.834,36 €

Tabella 10 - Conteggio codici con marca "BMW"

COD_DS	# in contratto_ven_nuovo	# in contratto_ven_usato	# fatture_att_testata	Valore economico
33		2026	392	5.578.435,36 €
80	7306		3575	159.308.805,70 €
MER		10673	1955	33.060.601,69 €

Tabella 11 - Conteggio codici con marca "MERCEDES"

COD_DS	# in contratto_ven_nuovo	# in contratto_ven_usato	# fatture_att_testata	Valore economico
VLK		3746	1392	15.602.379,88 €
101		1669	356	3.774.816,82 €
50	3839		332	8.019.818,92 €

Tabella 12 - Conteggio codici con marca "VOLKSWAGEN"

COD_DS	# in contratto_ven_nuovo	# in contratto_ven_usato	# fatture_att_testata	Valore economico
NSBOLH		3127	2	4.900,00 €
2222		7588	1433	28.992.885,00 €
MIN		4	707	11.625.743,00 €

Tabella 13 - Conteggio codici con marca "MINI"

COD_DS	# in contratto_ven_nuovo	# in contratto_ven_usato	# fatture_att_testata	Valore economico
SMA		4318	879	6.549.194,00 €
2259		253	42	366.045,00 €
79	3837		1757	29.068.347,00 €

Tabella 14 - Conteggio codici con marca "SMART"

COD_DS	# in contratto_ven_nuovo	# in contratto_ven_usato	# fatture_att_testata	Valore economico
OENVWX		69	14	66.105,00 €
8		1832	374	2.200.686,57 €
CFLVLD		1	1	300,00 €
FIA		2339	647	3.370.866,64 €

Tabella 15 - Conteggio codici con marca "FIAT"

COD_DS	# in contratto_ven_nuovo	# in contratto_ven_usato	# fatture_att_testata	Valore economico
144		545	121	2.134.869,62 €
72	28			
JEE		451	98	1.480.446,09 €

Tabella 16 - Conteggio codici con marca "JEEP"

Le tabelle sopra presentate mostrano come alcuni codici siano effettivamente utilizzati o per i contratti vendita di autoveicoli nuovi, o per contratti vendita di autoveicoli usati. Nei contratti vendita relativi all'usato, risultano però essere utilizzati più codici per indicare la stessa marca. Una possibile ipotesi sul perché questi codici siano così gestiti potrebbe riguardare la recente fusione che ha interessato il Dealer, oggetto di analisi, con un'altra azienda del settore e quindi, l'utilizzo di più codici potrebbe essere il risultato di un processo di integrazione tra le diverse modalità di gestione interne dei processi. Si è andati quindi ad analizzare questi codici in relazione al loro utilizzo temporale e geografico (in quali sedi).

Andando ad estrarre dalle tabelle le sedi in cui sono stati utilizzati ed il periodo temporale di utilizzo, non risulta esserci nessuna correlazione temporale o geografica con l'impiego di questi codici, quindi sembra che siano utilizzati scegliendo senza nessun criterio e l'ipotesi sopra definita non risulta valida. Una volta analizzato il campo descrizione dell'attributo, ed avendo ricercato errori che fossero attribuibili a problemi di semantica o a problemi legati all'errore umano di battitura, è stata spostata l'attenzione sulle relazioni riguardanti l'attributo della marca.

La seguente figura mostra i vincoli relazionali di chiave esterna che riguardano questo attributo.

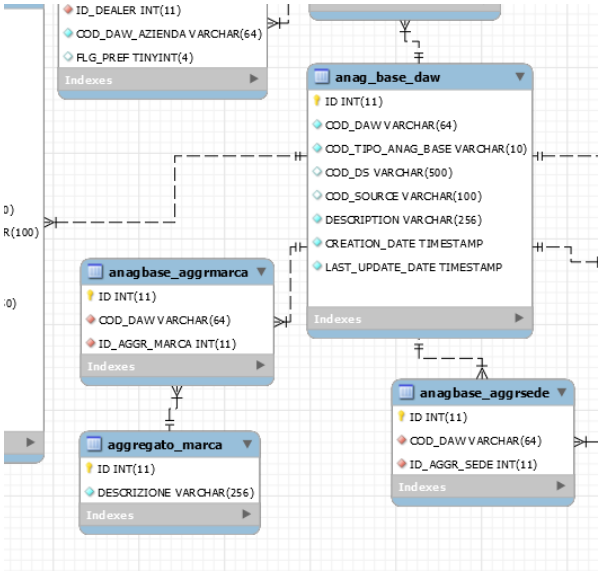

Figura 12 - Porzione di schema ER relativo all'attributo "Marca"

Dalla figura 12 è possibile vedere che, all'interno del database, esistono delle tabelle che raggruppano i diversi codici utilizzati per indicare il medesimo marchio: anagbase_aggrmarca e aggregato_marca. Secondo questo schema, le tabelle raggruppano i cod_daw[5], a cui viene associato un unico campo descrittivo. Questo codice, però, non risulta esser utilizzato all'interno delle tabelle relative ai contratti, alle fatture, agli ordini di lavoro o allo stock veicoli, in cui è presente il solo cod_ds e dove, quindi, il campo descrittivo risulta ancora errato e non aggregato.

4.4 ANALISI DELL'ENTITÀ SOGGETTO ALL'INTERNO DEL DATABASE

Come per l'entità veicolo, è stata effettuata un'analisi sull'entità riguardante il soggetto, chiamato anche Driver. Come già descritto precedentemente, in questa tabella si trovano tutte quelle persone che, in qualche modo, sono entrate in relazione con il

[5] Il cod_daw è un codice creato durante l'import dei dati e associato ad un cod_ds, codice utilizzato nel sistema operazionale del Dealer.

Dealer, sia solo per una proposta di offerta che per un contratto di acquisto vero e proprio o per un servizio di assistenza in officina.

Nei documenti riguardanti il tracciato dei dati dalle fonti alimentanti al repository centrale, in merito alle informazioni relative alle regole di duplicazione dati del soggetto, è possibile leggere che, dato che gli stessi possono provenire da più fonti (Sipad.X e LeadSpark), è stato definito un algoritmo per "approssimazione" per la determinazione del record univoco. Le prime informazioni che rendono un record univoco sono il Codice Fiscale/P.Iva. In assenza di queste informazioni viene verificato il valore di Telefono Principale e Email Principale. In assenza di queste informazioni viene verificato il valore di Cognome e Nome / Denominazione. Il valore NULL viene considerato come informazione non presente, pertanto in presenza di altri valori uguali, viene generato un unico record.

Sono state effettuate delle interrogazioni per controllare che non ci fossero errori di duplicazione soggetto nella relativa tabella. Entrambe le query restituiscono il conteggio dei record multipli, raggruppati prima per codice fiscale e poi per partita iva.

```
SELECT COUNT(*), COGNOME, NOME, UPPER(CODICE_FISCALE), PARTITA_IVA
FROM soggetto
GROUP BY UPPER(CODICE_FISCALE)
```

```
SELECT COUNT(*), COGNOME, NOME, CODICE_FISCALE, PARTITA_IVA
FROM soggetto
GROUP BY PARTITA_IVA
```

La prima interrogazione restituisce 353 record, riferiti a 353 soggetti, che risultano inseriti due volte poiché in una riga è valorizzato sia il campo relativo al codice fiscale che il campo relativo alla partita iva, mentre nell'altra riga risulta valorizzato solo il

campo codice fiscale, oppure il codice fiscale risulta essere inserito una volta con lettere maiuscole ed un'altra con lettere minuscole.

La seconda interrogazione, invece, restituisce 167 soggetti multipli, sempre seguendo la logica scritta sopra.

Si ricorda che sia il campo codice fiscale che la partita iva fanno parte della chiave primaria.

4.5 DEFINIZIONE DI INDICATORI DI QUALITÀ DEL DATO

A valle delle analisi effettuate sono stati definiti degli indicatori che misurassero l'impatto e la dimensione degli errori riconducibili alla qualità dei dati all'interno del database. Si immagina che questi indicatori verranno poi utilizzati per un monitoraggio interno della qualità del database e verrà quindi presentata una possibile dashboard.

Di seguito si elencano gli indicatori individuati e viene fornita una descrizione più amplia riguardo al metodo di calcolo e dell'importanza che essi assumono nel contesto descritto nella tesi.

Nome	**Tasso di errore del campo "marca"**
Descrizione	Percentuale di record con campo descrizione errato
Scala di misura	Rapporto
Unità di misura	-
Valore di riferimento	0%
Interpretazione e misura	Con questo indicatore si vuole calcolare la percentuale dei campi errati che riportano la descrizione della marca. Il campo viene confrontato con una tabella comparativa e così definito se è accettato o no a seconda che ricada nel range di valori definiti. V_{errore}= valore dell'indicatore M= totale dei record associati alla marca m= numero di marche errate $$V_{errore} = \frac{m}{M} * 100$$

Questo primo indicatore pone l'attenzione sul campo descrittivo relativo alla marca. Molti KPI utilizzati all'interno della Dashboard della Business Intelligence calcolano valori aggregandoli per marca, come ad esempio la composizione delle vendite o dello stock sulla base dei vari brand. Risulta quindi opportuno ripulire e rendere omogeneo questo campo, andando ad eliminare gli errori di battitura e correggendo tutti i valori di

marche non attribuibili a nessun brand. Ciò è fatto mediante una query di confronto con una seconda tabella contenente un elenco accurato di tutte le marche di autoveicoli esistenti. La query utilizzata è nella forma:

```
SELECT *
from DB1.Table

except

SELECT *
from DB2.Table
```

È opportuno effettuare ulteriori verifiche sui valori restituiti dalla query ed andare ad individuare puntualmente i record errati e da valutare per effettuare le modifiche necessarie.

Nome	**Codici marca multipli**
Descrizione	Percentuale
Scala di misura	Rapporto
Unità di misura	-
Valore di riferimento	0-3%
Interpretazione e misura	Questo indicatore indica la percentuale di marche in cui vengono utilizzati più codici per riferirsi al medesimo brand. V_{codice}= Valore dell'indicatore r= numero di marche con codice multiplo R= numero totale delle marche $$V_{codice} = \frac{r}{R} * 100$$

Anche questo indicatore, come il precedente, misura la qualità del dato in riferimento alla marca del veicolo. Si vanno ad identificare tutti quei record che presentano più codici distinti in relazione alla medesima marca e divisi per la totalità dei codici presenti. Come però evidenziato nel paragrafo 4.3, è emerso che il Dealer presenta più codici relativi alla medesima marca poiché questi codici vengono utilizzati in processi differenti, ed in particolare viene utilizzato un codice nelle operazioni di officina, un

codice per i contratti di vendita di auto nuove ed un codice per i contratti relativi alle auto usate. In riferimento a questa osservazione non si considerano errati quelle marche che presentano tre codici distinti per i tre processi, ma vengono considerate errate quelle marche che presentano più codici distinti in riferimento al medesimo processo.

Nome	**Unicità del soggetto**
Descrizione	Numero di soggetti duplicati
Scala di misura	Assoluta
Unità di misura	Unità
Valore di riferimento	0
Interpretazione e misura	L'indicatore individua il numero di righe multiple, presenti nella tabella soggetto, correlato al medesimo soggetto $V_{unicità}$= valore dell'indicatore S= numero di record con uguale Codice Fiscale o PIVA $$V_{unicità} = \sum S$$

A differenza dei precedenti indicatori, questo indicatore pone l'attenzione sulla qualità del dato in relazione all'entità soggetto. Dall'analisi effettuata precedentemente è emerso che ci sono molti record multipli, relativi al medesimo soggetto. Questi record sono stati trovati con le interrogazioni già scritte nel precedente paragrafo, ma che vengono riportate poiché utilizzate per estrarre i valori impiegati nel calcolo del KPI. LE interrogazioni sono le seguenti:

```
SELECT COUNT(*), COGNOME, NOME, UPPER(CODICE_FISCALE), PARTITA_IVA
FROM soggetto
GROUP BY UPPER(CODICE_FISCALE)
```

```
SELECT COUNT(*), COGNOME, NOME, CODICE_FISCALE, PARTITA_IVA
FROM soggetto
GROUP BY PARTITA_IVA
```

Nome	**Completezza record soggetto**
Descrizione	Numero di record con almeno uno dei seguenti campi nullo o non valido: CF, PIVA, numero di telefono, e-mail
Scala di misura	Rapporto
Unità di misura	-
Valore di riferimento	0
Interpretazione e misura	L'indicatore restituisce la percentuale di tutti quei record che presentano almeno un campo nullo o non valido negli attributi relativi a: Codice Fiscale e/o PIVA, numero di telefono ed e-mail. $V_{completezza}$= valore dell'indicatore d= Numero d record con almeno un campo nullo tra quelli definiti D= Numero di record totale $$V_{completezza} = \frac{d}{D} * 100$$

L'indicatore descritto sopra pone l'attenzione sul problema della completezza del record. Si considera completo quel record che presenta valori non nulli o non errati nei seguenti campi: Codice fiscale e/o Partita Iva, Numero di telefono ed e-mail. I primi due campi saranno quasi sempre compilati, poiché parte della chiave primaria. Si considerano altresì importanti i campi relativi al numero di telefono e alla mail del soggetto poiché queste informazioni vengono utilizzate per il calcolo del valore di conversione su cui si basa l'analisi dei Lead, o potenziale cliente. Per ogni dealer, l'obiettivo finale è migliorare i risultati di vendita e, a tal fine, viene tracciato ogni fase del processo, dai click del lead fino alla firma del contratto che fa di lui un cliente acquisito. È evidente che più i dati sono completi e accurati, minore è il margine di rischio di tali decisioni e maggiore il ritorno sull'investimento.

4.6 CALCOLO DEGLI INDICATORI E VISUALIZZAZIONE DASHBOARD

In questo paragrafo verranno presentati i risultati dei calcoli effettuati per determinare gli indicatori definiti precedentemente. In seguito, verrà proposta una dashboard grafica per una visualizzazione più intuitiva ed immediata dell'attuale situazione analizzata.

4.6.1 INDICATORE SUL TASSO DI ERRORE DEL CAMPO DESCRITTIVO RELATIVO ALLA MARCA

Le marche che risultano non corrette sono quelle già presentate in Tabella 4:

COD_DS	DESCRIPTION
FGGRFO	BARTOLETTI
SQWODK	BNW
CHA-M	CHATENET MOTO
CHC	CHEVROLET (K)
DRM	DR MOTOR
GAC	GAC GONOW GA200 2.0
BOLHPQ	GARAGE ITALIA
2715	GREAT WALL MOTOR
IMP	IMPORTAZIONE
LIG-M	LIGIER MOTO
99	MARCHIO GENERICO
VQHTC	MERCEDES V.C.
36	MERCEDES V.I.
766	MICROCAR
MIC-M	MICROCAR MOTO
78	MITSUBISHI FUSO/CANTER
100	PETTENELLA
PIA-M	PIAGGIO MOTO
PMF-M	PIAGGIO MOTOFURGONI
LWDGDR	SCANIA CV 480
CQJXFW	SCANIA CV R 420
MFOSRP	SCANIA R500
76	TRUCK - V.I.
77	VAN - V.C.
1	WOLKSVAGEN

Tabella 17 - Risultato marche non corrette

A valle di questi risultati va fatta un'analisi più dettagliata ed aggiungere alcune considerazioni. Se consideriamo i campi relativi a "TRUCK - V.I." e "VAN - V.C." va specificato che nonostante non siano delle marche vere e proprie, è giusto però considerarle come tali perché utilizzati solo nell'ambito delle vendite dell'usato e raggruppano tutta la categoria merceologica e non il particolare brand. Infatti, sia che si faccia riferimento alle logiche di calcolo delle provvigioni ai venditori, che alle logiche di calcolo dei KPI utilizzati nella Business Intelligence, non risulta rilevante, all'interno dei processi, scorporare questi campi e scendere ad un ulteriore livello di dettaglio.

Si mostrano i valori, che concorrono al calcolo dell'indicatore, ottenuti:

- Totale record associati alla marca: 206
- Numero di marche considerate errate: 23
- Valore dell'indicatore V_{errore}= 11%

4.6.2 INDICATORE SULLA MOLTEPLICITÀ DEI CODICI

Per ottenere il valore di questo indicatore sono state eseguite delle interrogazioni sulle tabelle relative ai contratti di vendita nuovo ed usato. Da queste interrogazioni è emerso che nella tabella dei contratti nuovi si fa sempre riferimento ad un unico codice relativo alla marca, e non vengono utilizzati codici differenti in riferimento alla medesima marca. Discorso non valido, invece, per l'analisi effettuata nella tabella relativa ai contratti di vendita dell'usato, in cui risulta evidente l'utilizzo di differenti codici in relazione alla stessa marca.

- Totale record associati alla marca: 206
- Numero di marche con codice multiplo: 55
- Valore dell'indicatore V_{codice}= 27%

4.6.3 INDICATORE SULL'UNICITÀ DEL SOGGETTO

Le interrogazioni per determinare il valore di questo indicatore sono state effettuate sulla tabella relativa al soggetto. Questa tabella ha una dimensione di circa 500.000 righe.

- Numero di record con pari codice fiscale: 353
- Numero di record con pari partita iva: 167
- Valore dell'indicatore $V_{\text{unicità}}$= 520

Dalle interrogazioni emerge che, in relazione al totale dei record presenti nella tabella, non sono molti i soggetti presenti in modo multiplo all'interno del database. Questo significa che sono rari quei casi in cui viene popolata erroneamente la tabella ed è un dato facilmente correggibile.

4.6.4 INDICATORE SULLA COMPLETEZZA DEL RECORD RELATIVO AL SOGGETTO

Come per l'indicatore precedente, i valori sono stati estratti dalla tabella soggetto.

- Numero di record relativi al soggetto: 461973
- Numero di record con almeno un campo nullo: 65743
- Valore dell'indicatore $V_{\text{incompletezza}}$= 14%

Per ciascun indicatore è possibile definire un periodo di tempo in cui ripetere le interrogazioni ed estrarre gli indicatori di interesse. Ciò mostrerebbe non solo una fotografia della situazione attuale, ma anche un'evoluzione temporale della qualità

dell'informazione all'interno del database. È così possibile riportare i risultati ottenuti in un cruscotto.

4.6.5 PRESENTAZIONE DELLA DASHBOARD

Contestualmente all'analisi effettuata e al calcolo degli indicatori, è possibile creare una dashboard in merito alla rappresentazione dei valori degli indicatori, per ottenere una visualizzazione immediata dell'attuale situazione qualitativa del database.

A partire dalle misure ottenute, è possibile ottenere dei grafici che mostrano sia una foto attuale della situazione, sia l'andamento temporale, così da vedere se ci sono stati miglioramenti o peggioramenti.

La figura sotto rappresenta il mockup del cruscotto, con i risultati degli indicatori.

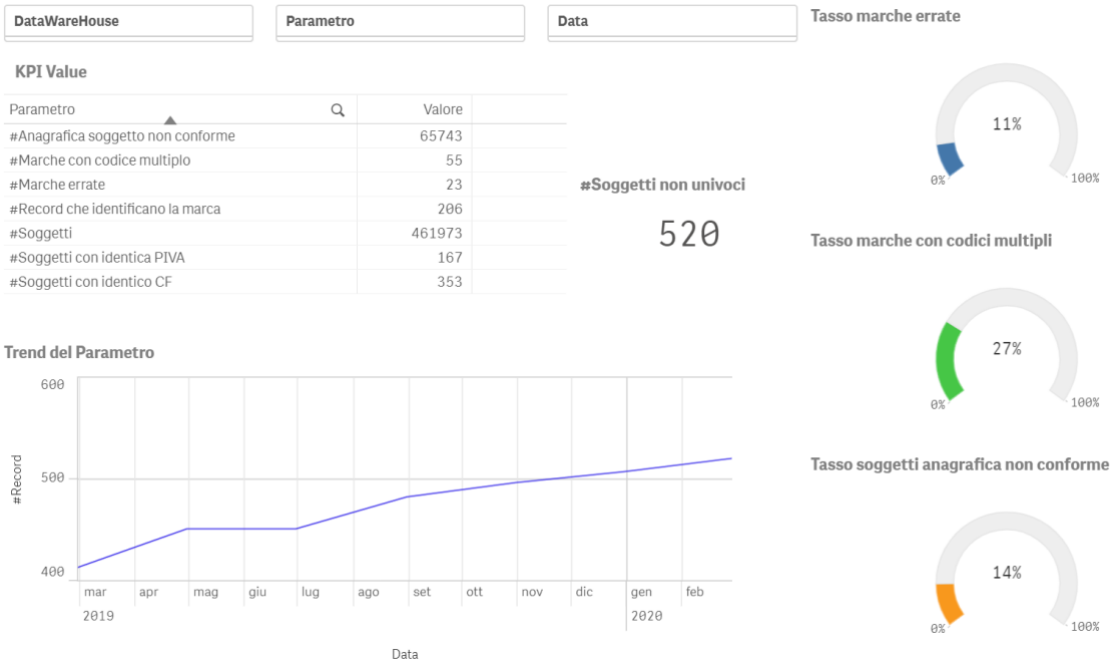

Figura 13 - Mockup Dashboard

Il cruscotto è costituito da quattro grafici: nella parte destra del cruscotto sono riportati i valori degli indicatori, anche rappresentati tramite un grafico gauge, noto anche come grafico a tachimetro. Gli indicatori rappresentati tramite grafico a tachimetro sono: indicatore del tasso di errore del campo marco, l'indicatore relativo ai codici marca multipli e l'indicatore che misura la completezza del record soggetto. Nella parte bassa della dashboard è presente un istogramma che riporta il trend dell'indicatore, selezionabile tramite menù a tendina nella parte alta della dashboard. In alto a sinistra vengono riportati tutti i valori che concorrono al calcolo degli indicatori, così da poter esaminare i valori ottenuti ed individuare la possibile area di errore.

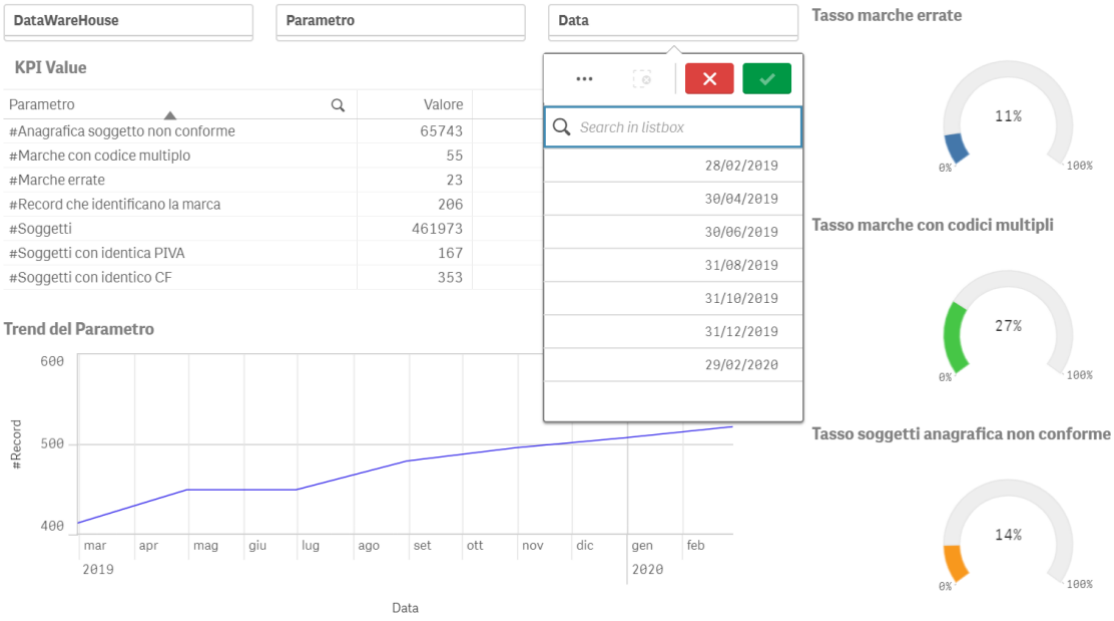

Figura 14 - Selezione della data sulla dashboard

CONCLUSIONI E SVILUPPI FUTURI

Il settore automotive in Europa e in Italia è oggigiorno soggetto a importanti cambiamenti, dati dalle continue innovazioni in ambito digitale e da un mutamento continuo della percezione dei clienti sia per i prodotti, ma soprattutto per i servizi offerti in questo settore. Questa trasformazione non solo ha impattato, ed impatta, i grandi produttori di auto, ma anche gli attori di tutta la filiera. In particolar modo i dealer italiani, il cui business è ancora impostato molto a livello familiare, stanno rivalutando le modalità di contatto con il cliente, non affidandosi solo al venditore in concessionaria, che ricopre comunque un ruolo decisivo in fase di vendita, ma anche ai servizi digitali connessi. Nonostante stiano ancora soffrendo della scarsa digitalizzazione dei processi gestionali, è possibile osservare un graduale avvicinamento ed utilizzo di mezzi sempre più omnicanale, come per esempio l'utilizzo della firma elettronica per la stipula de contratti di vendita o dei finanziamenti e l'impiego di App mobile per proporre al cliente la prenotazione del servizio di officina o per poter ricercare l'auto di interesse all'interno dello stock veicoli.

Questa trasformazione digitale dei processi, però, non solo fornisce un valido strumento per lo sviluppo di tutto il business, ma crea un indotto di dati sempre maggiore. L'origine di questi dati è molteplice, andando così a creare disomogeneità nell'informazione che risiede nei sistemi centrali.

In seguito alla progettazione di un repository centrale, in cui confluiscono tutti i dati provenienti dai processi quotidiani del dealer, è emersa questa disuguaglianza di informazione tra le entità definite, spesso accompagnata da bassi livelli di accuratezza. L'importanza dell'alta qualità dei dati è dettata dalle esigenze delle attività operative e analitiche che li processano.

Ciò che emerge da questo studio è la necessità di uno strumento di analisi e monitoraggio della qualità dei dati all'interno del database, in modo da poter meglio gestire e renderli fruibili per le attività di business e la misurazione delle performance. Gli indicatori individuati e la dashboard creata sono un valido strumento per il processo di data governance, così da poter identificare i possibili problemi, garantire un livello accettabile di fiducia dei dati e soddisfare le esigenze aziendali dell'organizzazione.

Gli sviluppi futuri possono essere ipotizzati lungo due direttive: da un lato l'automatizzazione dei processi di monitoraggio della qualità del dato in sistemi di data integration in produzione; da un altro lato la costruzione di un toolset di strumenti e metodi per effettuare misurazioni sui sistemi origine con lo scopo di stimare con buona accuratezza l'efficacia ed i costi di integrare i sistemi stessi.

Nel primo dei due casi, l'evoluzione consisterà nel rendere automatico tutto il processo di analisi e reporting, perché, se è necessario identificare e misurare degli indicatori di qualità, è quanto mai opportuno integrare le misure in un processo che controlli la conformità dei dati senza il diretto intervento umano, faccia dei confronti con le soglie di accettabilità ed invii degli alert agli amministratori in modo da intraprendere azioni specifiche quando necessario.

Nel secondo, l'evoluzione porterà alla formalizzazione di best practices e strumenti informatici a supporto, con lo scopo di ottenere indicazioni quantitative sulla qualità del dato nel sistema origine prima di effettuare le operazioni di data integration e in modo da ridurre i rischi e le incertezze nell'implementazione delle procedure di data integration.

BIBLIOGRAFIA E SITOGRAFIA

[1] Italia Bilanci (2018), Automotive Dealer Report

[2] Roland Berger (2012), 21st Century Car Distribution in Europe, MBA Project IESE

[3] McKinsey&Company (2016), Automotive revolution – perspective towards 2030. How the convergence of disruptive technology-driven trends could transform the auto industry.

[4] Accenture (2018), The new automotive dealer: designed for me

[5] Acea (2016), The 2030 Urban Mobility Challenge, Acea Contribution

[6] Batini Carlo, Scannapieco Monica (2006), Qualità dei Dati Concetti, Metodi e Tecniche

[7] Thomas C. Redman (1996), Data quality for the information age, Artech House

[8] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang (2002), Data Quality Assessment

[9] Rahm, E. and Do, H. H. (2000), Data Cleaning: Problems and Current Approaches". IEEE Bulletin of the Technical Committee on Data Engineering

[10] Golfarelli, Rizzi (2006), Data Warehouse, teoria e pratica della progettazione

[11] https://www.eng.it/

[12] http://www.oica.net/

[13] www.anfia.it

INDICE DELLE FIGURE

INDICE DELLE TABELLE