POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Gestionale
A.A. 2022/2023
Sessione di Laurea Luglio 2023

Tesi di Laurea Magistrale

Analisi e ottimizzazione dell'apertura degli store nel GDO

Relatori

prof.ssa COLOMBELLI ALESSANDRA

Candidato

Alice Salanitri

Alla mia famiglia

Sommario

Indice delle figure

INTRODUZIONE

Il presente elaborato è il risultato dell'esperienza di tirocinio svoltasi tra i mesi di marzo e luglio 2023 presso la sede torinese dell'azienda Mediamente Consulting s.r.l. .

Fondata nel 2012, Mediamente Consulting è una società di consulenza informatica specializzata in sistemi di supporto alle decisioni. Al suo interno sono presenti cinque diverse aree di competenza: Data Integration, Corporate Performance Management, Advanced Analytics, Business Intelligence e Infrastruttura Tecnologica. La varietà di tali competenze offre ai clienti la possibilità di beneficiare di un'infrastruttura tecnologica avanzata in grado di gestire e analizzare dati complessi , fornendo loro informazioni essenziali per le decisioni strategiche aziendali e garantendo un supporto costante durante tale processo decisionale.

Viviamo in un'epoca caratterizzata dalla digitalizzazione della società. Negli ultimi anni, si è passati da una realtà prevalentemente analogica ad un ambiente digitale interconnesso, grazie alla rapida evoluzione della tecnologia e dell'informatica che ha avvicinato sempre di più uomo e "macchine".

È proprio in tale contesto che i dati hanno subito una crescita esponenziale: smartphone, social media, e-commerce hanno man mano portato ad un incremento considerevole della quantità di dati generati quotidianamente, divenendo sempre più importanti per le decisioni strategiche.

Oggi, quasi tutti i settori si servono dei dati per prendere decisioni: le aziende, ad esempio, si affidano all'analisi dei dati per comprendere il comportamento dei consumatori e dei competitor, creando campagne marketing mirate ad ogni cluster di consumatori; oppure, nel settore sanitario l'analisi può aiutare la ricerca, permettendo di migliorare la prevenzione e la cura delle malattie.

Tuttavia, per far sì che i dati creino un vero e proprio vantaggio competitivo è indispensabile investire costantemente in strumenti all'avanguardia in grado di gestire l'immensa mole di informazioni creata. Negli anni, le architetture di gestione dei dati si sono via via evolute, passando dal tradizionale data warehousing ad architetture sempre più complesse che consentono di elaborare, anche in tempo reale, dati di formati diversi (sia strutturati che non), supportando transazioni ad alta velocità senza compromettere le prestazioni ottimali.

CASO DI STUDIO

Per svolgere il mio progetto di tesi il cliente assegnatomi da Mediamente Consulting è un player della grande distribuzione organizzata, facente parte di un gruppo che controlla più società operanti in differenti aree di business, dal campo immobiliare a quello produttivo. Nel seguente elaborato ci si focalizzerà esclusivamente all'area legata alla grande distribuzione organizzata.

La GDO in esame presenta un'area di competenza concentrata principalmente nel Centro-Sud d'Italia, con una maggiore diffusione nelle regioni della Basilicata e della Campania. Comprende quattro tipologie di brand, tre facenti parte del settore B2C e uno del settore B2B. I quattro brand si distinguono per il loro format distributivo, ovvero in base a diverse specifiche tecniche (quali la dimensione del punto vendita, l'assortimento dei prodotti, la posizione, le politiche di prezzo) che ne definiscono in modo tecnico l'appartenenza a una delle categorie di GDO (come ad esempio discount, ipermercato, cash & carry e così via).

- Brand 1: comprende i *punti vendita di traffico*, ovvero i maxi e iper-store la cui superficie va dai 1000 fino ai 4000 mq. Offrono un ampio assortimento di prodotti. È riconosciuto come il "supermercato per tutta la famiglia" , in grado di soddisfare le esigenze di ogni consumatore. Il cliente tipico è colui che effettua una spesa di grandi dimensioni.
- Brand 2: caratterizzato invece dai *punti vendita di prossimità* con store fino a 800 mq. Soddisfa i bisogni dei clienti con elevata frequenza d'acquisto. È visto come il "supermercato sotto casa", comodo per la spesa giornaliera.
- Brand 3: è centrato sui formati convenienza, ideale per chi cerca comodità, qualità e convenienza tutti i giorni.
- Brand 4: fa parte del settore professional con la presenza di cash & carry tra i 2000 e i 3000 mq.

In questa tesi, l'analisi si concentrerà esclusivamente sul settore consumer dell'azienda; quindi, non verrà preso in considerazioni il brand 4.

OBIETTIVO

In un contesto altamente competitivo come quello della Grande Distribuzione Organizzata l'impiego di strumenti di analisi di dati risulta essere essenziale per garantire una costante espansione.

Attraverso l'implementazione di una processo ETL si intende creare un Data Mart che permetta di integrare i dati aziendali con i dati provenienti da fonti open, fornendo così una visione più ampia

del contesto in cui opera il cliente. L'obiettivo della presente tesi consiste nell'individuare le migliori localizzazioni per l'apertura di nuovi punti vendita. Riuscire a riconoscere le zone più promettenti e profittevoli per aprire nuovi store porta numerosi vantaggi al cliente, tra cui:

1. Riduzione della concorrenza: uno studio preliminare della zona in cui si intende aprire un punto vendita consente di individuare le aree in cui la concorrenza è meno agguerrita, permettendo così all'azienda di posizionarsi in mercati meno saturi. In questo modo aumentano le probabilità di successo e si riduce la pressione competitiva.
2. Personalizzazione delle offerte: avendo una maggiore informazione su comportamenti della popolazione e sulle caratteristiche dei luoghi in questione l'azienda può personalizzare le offerte al fine di soddisfare le esigenze dei clienti locali, migliorando così la customer experience.
3. Maggiori economie di scala: aprire punti vendita in zone ottimali consente all'azienda di beneficiare di economie di scala, riuscendo così a diminuire i costi operativi e ad aumentare le vendite.
4. Migliori negoziazioni con i fornitori: l'apertura in luoghi strategici non comporta benefici solo per l'azienda, ma anche per i suoi fornitori. Infatti, la presenza di punti di vendita in zone privilegiate assicura un maggiore afflusso di clientela, offrendo una maggiore visibilità ai vari fornitori. Questa situazione favorisce la creazione di partnership , che consentono all'azienda di richiedere offerte e sconti vantaggiosi. Ciò si traduce in prezzi competitivi per i clienti, attraendo così un numero sempre maggiore di acquirenti.

Di seguito verranno presentati brevemente i capitoli che compongono l'elaborato.

Il primo capitolo introduce lo stato dell'arte economico nel settore della grande distribuzione organizzata (GDO). Esso offre una panoramica generale sulle caratteristiche distintive delle imprese della GDO e sugli obiettivi che queste si propongono di raggiungere. Vengono successivamente analizzate le strategie e le tecniche di posizionamento adottate dalle aziende operanti in questo settore.

Il secondo capitolo si focalizza maggiormente sugli aspetti tecnici, trattando argomenti quali la Business Intelligence e i Big Data. Dopo una introduzione preliminare, vengono spiegati e descritti i numerosi vantaggi derivanti dall'utilizzo di tali strumenti. Una particolare attenzione è dedicata alla gestione e all'organizzazione dei dati, andando ad approfondire concetti quali Datawarehouse, Data Lake e i modelli di organizzazione dati come lo Star Schema e lo Snowflake. Infine, viene introdotto il processo ETL (estrazione, trasformazione e caricamento dei dati), che verrà ulteriormente approfondito nel capitolo successivo.

Il terzo capitolo è dedicato alla descrizione delle analisi condotte e alla metodologia utilizzata per la raccolta dei dati. Nel dettaglio, verrà prima spiegato il processo ETL aziendale, ottenendo come risultato lo Star schema. Successivamente, verrà svolta la cross analysis tra i dati interni all'azienda e dati provenienti da fonti aperte al fine di individuare le provincie più efficaci per l'apertura di nuovi punti vendita.

Infine, nel capitolo conclusivo viene affrontato il tema della Data Visualization, sottolineando l'importanza che questa riveste nell'analisi dei dati. Tra i diversi strumenti di Data Visualization è stato deciso di utilizzare Power BI. Dopo una dettagliata descrizione di questo strumento sono stati presentati alcuni dei risultati ottenuti dalle analisi precedentemente effettuate al fine di mostrare concretamente come la Data Visualization possa trasmettere efficacemente le informazioni ricavate.

1. STATO DELL'ARTE ECONOMICO

1.1 Grande distribuzione organizzata: principali caratteristiche

La Grande Distribuzione Organizzata, conosciuta anche con l'acronimo GDO, è un sistema di vendita al dettaglio di prodotti di largo consumo, sia alimentari che non alimentari. Questi vengono venduti attraverso una rete di punti vendita, gestiti a libero servizio, e solitamente affiliati ad un gruppo o un'organizzazione comune che fornisce loro un supporto logistico, promozionale e operativo. (Tieri & Gamba, 2009)

In Italia, questa tipologia di distribuzione iniziò a svilupparsi a partire dagli anni Settanta del Novecento per rispondere alle nuove esigenze dei consumatori. Fino a quel momento, le attività commerciali venivano svolte esclusivamente nei piccoli punti vendita a conduzione familiare, che offrivano un limitato assortimento di prodotti. (Sbrana & Gandolfo, 2007)

Con l'avvento delle GDO, l'area di vendita è stata invece rinnovata: vengono prediletti punti con grandi superfici e con una ampia scelta di prodotti e nuovi format distributivi.

Un importante cambiamento viene apportato soprattutto dal punto di vista gestionale. Vengono distinte due macrocategorie di imprese, che si differenziano in base all'organizzazione interna (Report Doxee, 2020) :

- Distribuzione organizzata (DO): Ovvero una distribuzione che coinvolge esercenti di piccole e medie dimensioni, i quali collaborano mediante la costituzione di gruppi di acquisto per ottenere benefici contrattuali ma che rimangono giuridicamente indipendenti. Si trovano maggiormente nelle aree cittadine.
- Grande distribuzione (GD): insieme di punti di vendita gestiti da un unico soggetto proprietario o gruppo societario di imprese (gestione centralizzata). Gli store appartenenti a questa tipologia sono di grande dimensione e solitamente localizzati in zone periferiche. Questa organizzazione porta a numerosi vantaggi economici, in particolare si ha una riduzione dei costi grazie alle economie di scala nel comparto distributivo.

Tuttavia, la distinzione tra queste modalità gestionali diventa meno evidente nel momento in cui le imprese della GD concedono una maggiore autonomia ai singoli punti di vendita e quando i consorzi delle DO evolvono in forme più orientate al profitto e alla competitività nel mercato. (Tieri & Gamba, 2009)

1.2 Diversi format distributivi

Nella realtà italiana gli esercizi della GDO vengono classificati secondo diverse tipologie di format distributivi. Generalmente le varie configurazioni si distinguono sulla base della classe dimensionale della superficie di vendita e su diversi parametri quali: servizi offerti, livello di assortimento, prezzi e così via.

Principalmente i canali di vendita vengono distinti in: ipermercati, supermercati, liberi servizi, discount e cash and carry. (Panza, 2013)

1.2.1 IPERMERCATI

Sono esercizi con una superficie di vendita maggiore o uguale a 2500 mq. Grazie al grande spazio a disposizione propongono un assortimento ampio e profondo garantendo al consumatore una vasta scelta di prodotti. È generalmente realizzato nelle zone periferiche o suburbane, fuori dai centri abitati. Attrae clienti che si recano nello store per una spesa settimanale o maggiore. (NIELSEN, 2008)

1.2.2 SUPERMERCATI

Operano anch'essi nel campo alimentare ma dispongono di una superficie compresa tra 400 e 2499 mq. Hanno un vasto assortimento di prodotti, ma inferiore rispetto a quello degli ipermercati. Non possiede un target specifico di clientela, interessa principalmente i consumatori che vogliono risparmiare sia tempo che denaro e preferiscono fare la spesa presso strutture più facili da raggiungere. (NIELSEN, 2008), (Sbrana & Gandolfo, 2007)

1.2.3 LIBERI SERVIZI

Presenta una superficie compresa tra i 100 e 399 mq. Sono situati nei pressi di abitazioni e uffici e dispongono di un assortimento limitato di prodotti, per lo più quelli essenziali per la spesa alimentare. La clientela è composta principalmente da coloro che abitano nei pressi del punto vendita e che vi si recano per effettuare acquisti finalizzati ad integrare la spesa settimanale. (NIELSEN, 2008)

1.2.4 DISCOUNT

Questa categoria, contrariamente alle precedenti, non viene identificata in base alla dimensione ma si distingue dalle altre per le caratteristiche fisiche dei punti di vendita. Infatti, dispone di un allestimento più essenziale e di un assortimento prevalentemente unbranded. Si contraddistinguono

per i prezzi inferiori rispetto agli altri esercizi. I succitati punti vendita vengono classificati in hard e soft discount. I primi seguono una politica di riduzione dei prezzi e presentano prodotti senza marche, i secondi, invece, cercano di soddisfare maggiormente le esigenze dei consumatori proponendo anche prodotti freschi e di note marche commerciali (NIELSEN, 2008)

1.2.5 CASH AND CARRY

Punti di vendita specializzati nella vendita all'ingrosso, che viene effettuata esclusivamente da clienti in possesso di partita IVA. Non viene fornita assistenza alle vendite.

1.3 Obiettivi di espansione

Le aziende della grande distribuzione organizzata cercano costantemente nuove opportunità di crescita e strategie efficaci per il raggiungimento dei loro obiettivi.

È fondamentale adottare una strategia mirata e ben pianificata che guidi le decisioni e le azioni dell'azienda e che permetta di tradurre gli obiettivi in risultati concreti.

La strategia viene definita da Thompson come "il piano d'azione elaborato dal management per la gestione delle operazioni e delle attività di business dell'impresa". (Thompson, 2009)

È un piano a lungo termine che prevede la definizione di tutti gli obiettivi che l'azienda si prefissa di raggiungere e in cui vengono pianificati impegni, azioni e risorse necessarie al raggiungimento di questi. Deve poter essere mutevole e flessibile, in modo da riuscire a adattarsi a tutti i cambiamenti che possono sopraggiungere tanto internamente quanto esternamente all'impresa. Inoltre, un aspetto fondamentale nella scelta è cercare di distinguersi dai competitor, proponendo alla clientela prodotti, servizi nuovi o migliori rispetto a quelli già esistenti, al fine di ottenere un vantaggio competitivo. (Porter, 1985)

Obiettivo comune di tutte le aziende è quindi la creazione del valore per tutti gli attori coinvolti nel sistema, valore inteso come capacità dell'impresa di crescere e perdurare nel tempo, procurando anche benefici economici a tutte le parti interessate.

Nel caso delle GDO la crescita è strettamente legata al raggiungimento di una maggiore copertura geografica, che permette di ampliare la base di clienti e di catturare nuovi segmenti di mercato, ottenendo così una maggiore affluenza nei punti di vendita e un aumento del fatturato.

Al fine di ottenere tali risultati vengono progettate e pianificate diverse strategie, tra le più comuni troviamo:

- Apertura di nuovi punti di vendita: l'apertura di sedi in aree geografiche nuove o zone maggiormente strategiche consente il raggiungimento di un pubblico sempre più vasto. Questa decisione deve essere però ponderata, poiché numerosi sono i fattori che influenzano tale scelta.

Bisogna, in primo luogo, fare una analisi preliminare del mercato, analizzandone la domanda, valutando la presenza dei competitor e studiando le preferenze di consumatori. Non meno importante è lo studio della zona sulla quale si intende investire - valutandone le caratteristiche demografiche, economiche quali ad esempio numero di residenti, percentuale di povertà, tasso di disoccupazione e così via.

- Acquisizioni e fusioni: è il modo relativamente più semplice per espandersi e raggiungere gli obiettivi di crescita. Le acquisizioni, come dice il termine stesso, hanno lo scopo di acquisire il controllo di un'altra società. Questo permette un accesso più facile a mercati in cui, senza una base già esistente all'interno, risulterebbe molto complesso l'inserimento. La fusione, invece, consiste nell'unione di più imprese che creano un nuovo soggetto giuridico autonomo. (Fiori & Tiscini, 2014) . Queste due operazioni consentono alle aziende non solo di aumentare rapidamente la loro presenza nel mercato, ma anche di poter beneficiare delle varie sinergie operative ed economiche.
- Espansione internazionale: spesso diverse catene di GDO scelgono si espandersi anche fuori dalla propria nazione. I vantaggi derivanti da questa decisione sono molteplici, tra cui l'attrazione di nuovi consumatori (nella misura in cui si riesce a raggiungere una clientela sempre più diversificata), la creazione di economie di scala e un aumento dell'influenza sul mercato. Tuttavia, l'espansione internazionale può risultare molto complessa a causa delle ingenti spese necessarie per la realizzazione, per la difficile comprensione del mercato e delle normative vigenti. È quindi importante valutare attentamente la fattibilità del progetto, approfondendo ogni aspetto giuridico e finanziario. [1]
- Diversificazione dei format distributivi: un altro modo per attirare nuovi consumatori e aumentare le opportunità di vendita è quello di adattare i vari punti di vendita alle esigenze di ciascun segmento di clientela. Alcune aziende di GDO, ad esempio, offrono prodotti non solo di generi alimentari, ma anche reparti di abbigliamento, di elettronica o servizi di ristorazione, lavanderia.
- E-commerce: negli ultimi anni, i canali d'acquisto online si sono pian piano consolidati e numerosi consumatori usufruiscono di tali servizi per ordinare prodotti alimentari. Il Covid 19 ha avuto un impatto significativo sulle abitudini dei consumatori che, per paura di frequentare posti affollati, si sono dimostrati molto inclini all'acquisto tramite canali online.

Nonostante nell'ultimo anno la paura nei confronti del virus sia diminuita, studi dimostrano che, rispetto al 2020, il 9% in più dei consumatori è tuttora disposto a fare acquisti online [2]. È per questo motivo che l'e-commerce è diventato sempre più rilevante nella grande distribuzione organizzata. Questa strategia aiuta ad ampliare la portata geografica senza la necessità di dover aprire punti vendita fisici.

- Personalizzazione e marketing mirato: la possibilità di poter analizzare i dati dei clienti permette alle aziende di personalizzare le offerte e le promozioni, migliorando così la fidelizzazione degli acquirenti. Può divenire di grande vantaggio per le GDO l'utilizzo di attività di marketing digitale rispetto a quello tradizionale basato su cartelloni pubblicitari e volantini. Aiuta ad esempio a differenziare il proprio marchio da quello della concorrenza costruendo una relazione con il pubblico interessato; consente di comprendere le tendenze e le intuizioni dei consumatori; infatti, attraverso strumenti di analisi dei dati le aziende hanno differenti metriche di misurazione come, ad esempio, il numero di clic in un annuncio digitale. [3]

1.4 Tecniche di posizionamento

Dopo aver approfondito le varie strategie che vengono adottate dalle imprese per la loro crescita ed espansione, esaminiamo adesso le diverse metodologie attraverso cui l'azienda cerca di raggiungere tali obiettivi. In particolare, lo studio si concentrerà sulle tecniche utili all'identificazione delle migliori aree o mercati in cui aprire un nuovo punto vendita al fine di massimizzare il fatturato e soddisfare le esigenze dei clienti.

1.4.1 ANALISI DEL SETTORE

Prima di investire sull'apertura di nuovi punti vendita è di cruciale importanza fare una analisi preliminare del settore in cui si opera. Questa analisi serve per individuare e comprendere tutte le dinamiche ambientali che possono influenzare le performance delle imprese, permettendo di conoscere competitor e i loro possibili comportamenti. Diversi sono gli strumenti che ci permettono di attuare tale studio, tra questi troviamo il **modello delle cinque forze di Porter**. Questo mira ad individuare tutte le variabili che influenzano la competitività all'interno del settore includendo, oltre alle imprese direttamente concorrenti sul mercato, anche altri attori quali concorrenti potenziali, produttori beni sostitutivi, fornitori e clienti. (Ferrero, 2018). Risulterà così possibile valutare il potenziale successo di un nuovo punto vendita.

Per semplicità le cinque forze vengono suddivise in forze orizzontali (comprendenti le minacce di nuovi entranti, rivalità tra le imprese esistenti, minaccia di prodotti o servizi competitivi) e forze verticali (potere contrattuale dei fornitori e potere contrattuale degli acquirenti).

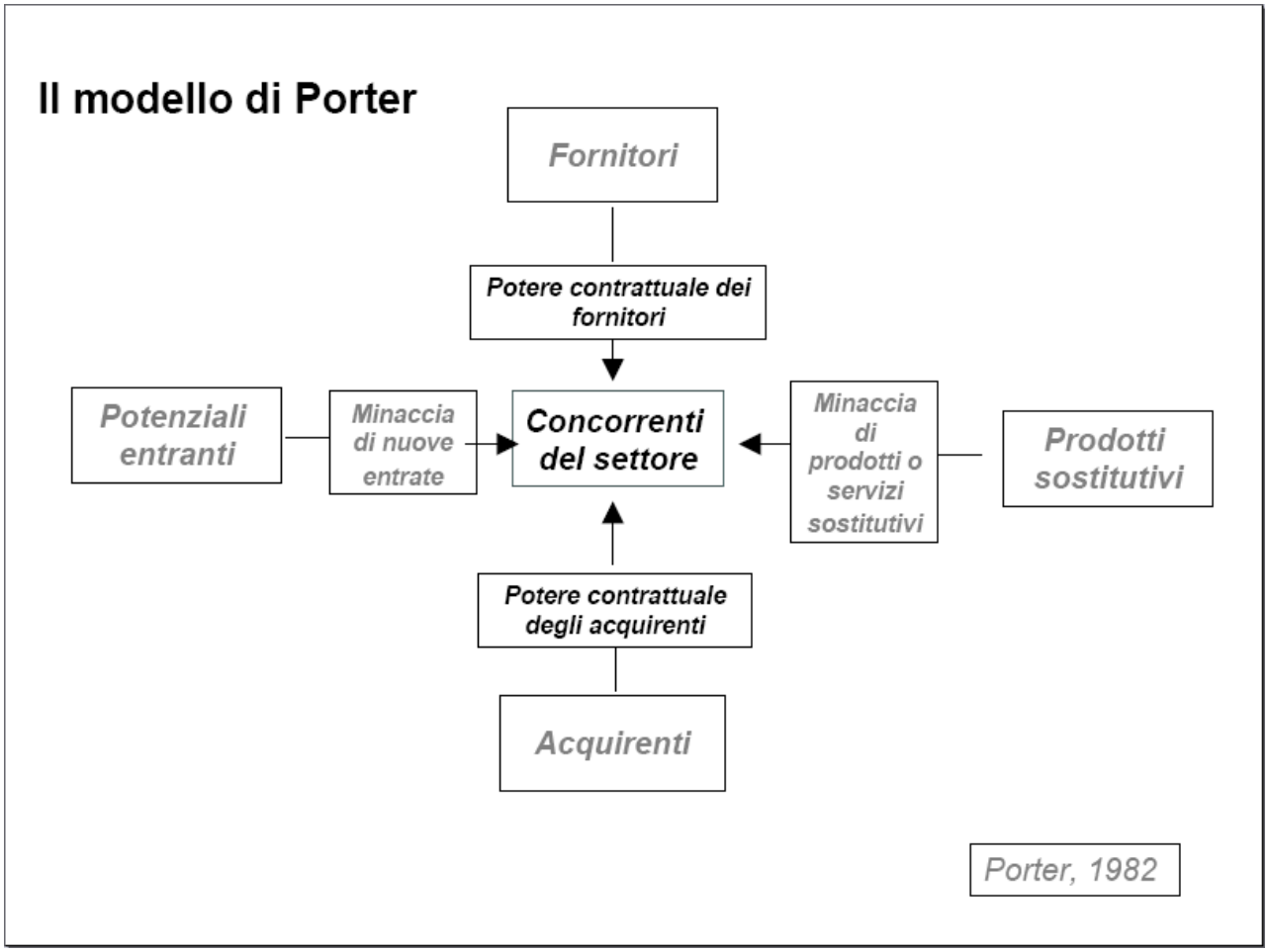

Figura 1: Modello delle 5 forze di Porter

Nello scenario delle GDO italiane è importante sottolineare quanto la concorrenza, intesa anche tra format distributivi differenti, risulti, a causa dell'elevato numero di imprese presenti nel mercato, frammentata. Questa forte competizione deriva principalmente da un' offerta poco differenziata che causa una sfida basata essenzialmente sui prezzi. In un contesto così altamente competitivo diventa fondamentale per le imprese adottare strategie di differenziazione, offrendo prodotti e servizi nuovi rispetto ai competitor al fine di attirare e fidelizzare sempre più clienti; devono, inoltre, innovarsi costantemente , investire in ricerca e sviluppo e adottare tecnologie innovative, quali analisi dei dati o l'intelligenza artificiale, così da essere sempre in grado di rispondere a tutte le esigenze dei consumatori.

Per quanto riguarda i produttori di beni sostitutivi, invece, dobbiamo prendere in considerazione coloro che offrono prodotti simili attraverso servizi differenti. La minaccia dei beni sostitutivi dipende da due fattori:

- Propensione dell'acquirente a sostituire il bene, che dipende a sua volta anche dai cosiddetti costi di switching (costi che il consumatore deve sostenere se decide di passare da un'azienda/prodotto/ servizio ad un altro);

- Dal rapporto qualità/prezzo. Infatti, più è semplice per il consumatore fare un confronto della qualità e dei prezzi di vari prodotti e servizi più elevata è la minaccia di sostituire quel prodotto.

Per poter gestire queste minacce le GDO devono attuare diverse strategie, quali, ad esempio, sviluppare servizi difficili da replicare così da non essere copiati dai competitor, oppure introdurre strumenti innovativi che migliorano l'esperienza del consumatore durante l'acquisto.
Un esempio di beni sostituti nel contesto della grande distribuzione organizzata è il "click and drive", in cui il consumatore sceglie online i prodotti da acquistare per poi recarsi di persona nel punto vendita e ritirare la spesa, o i servizi di "home delivery", in cui il consumatore sceglie online i prodotti che verranno consegnati direttamente a casa, grazie ad un personal shopper.
In questo caso, per rispondere all'incrementale utilizzo da parte dei consumatori dei servizi di e-commerce, negli ultimi anni le GDO stesse hanno iniziato ad offrire questi strumenti in modo da fornire un'alternativa conveniente all'acquisto tradizionale in negozio.
Quando si analizzano le forze verticali, è importante sottolineare che il potere contrattuale dei fornitori e dei consumatori può avere un impatto significativo sulla profittabilità dell'azienda. Infatti, più è elevato questo potere più questo ha un impatto negativo sull'azienda. Ad esempio, un potere maggiore dei fornitori consente loro di impostare prezzi più alti e di fornire prodotti a condizioni meno favorevoli; mentre, un elevato potere da parte dei consumatori permette di esercitare pressioni sui prezzi, richiedere servizi più vantaggiosi o minacciare di passare ai competitor.
Nel settore delle GDO, il potere contrattuale dei consumatori può essere considerato relativamente debole. Infatti, poiché i clienti finali sono generalmente singoli individui che effettuano acquisti di dimensioni ridotte rispetto al fatturato totale delle imprese, hanno limitate possibilità di negoziazione diretta con i rivenditori. Inoltre, il ruolo di intermediario della GDO tra consumatore e produttore riduce ulteriormente il potere contrattuale del cliente.
Tuttavia, non bisogna sottovalutare i vari aspetti che possono variare il loro potere di acquisto, come ad esempio cambiamenti dello stile di vita o cambiamenti delle condizioni della società (Colarieti, 2020) (Tieri & Gamba, 2009). Pertanto, è importante per le GDO adottare strategie mirate a migliorare la soddisfazione del cliente e a fidelizzarlo. A tal proposito, uno strumento efficace per raggiungere tale obiettivo è l'utilizzo dei dati; attraverso l'analisi dei dati relativi sia al comportamento che alle preferenze d'acquisto degli acquirenti, nonché alle loro abitudini e stili di vita, le imprese sono in grado di offrire offerte e servizi su misura per soddisfare le esigenze individuali.

Nel contesto delle GDO non meno importante è il potere contrattuale dei fornitori. Essi, infatti, possono influenzare le condizioni di fornitura, prezzi e termini contrattuali. Per poter gestire queste relazioni le GDO devono adottare differenti strategie. Innanzitutto, è importante non focalizzarsi solo su un unico fornitore, ma cercare di lavorare con una ampia gamma di fornitori differenti, così da evitare di essere troppo dipendenti da uno in particolare. È, inoltre, importante cercare di stringere una partnership con i diversi fornitori così da poter negoziare con loro. Ad esempio, se l'azienda riesce a determinare le migliori posizioni per l'apertura di nuovi punti di vendita attraverso uno studio approfondito di dati interni ed esterni all'azienda stessa, può ottenere un grande vantaggio competitivo anche rispetto ai fornitori. Individuare le zone e le località ottimali attirerà una clientela sempre più numerosa. Di conseguenza, offrendogli una maggiore visibilità e accesso a una clientela più vasta , la GDO diventa un canale di distribuzione attraente per loro. Ciò porta i fornitori ad essere più inclini a negoziare i prezzi e offrire condizioni contrattuali favorevoli per entrambe le patri coinvolte.

1.4.2 ANALISI AMBIENTE ESTERNO

Avere una panoramica dell'ambiente esterno può essere utile per le decisioni di posizionamento dei nuovi punti di vendita. Uno dei metodi usati per questo tipo di studio è rappresentato dalla PEST analysis.

Nonostante non si possa considerare una tecnica di posizionamento vera e propria, rappresenta comunque un utile strumento per esaminare tutti gli aspetti macro-ambientali esterni che possono avere un impatto sulle performance di un'azienda. Le variabili dell'analisi PEST includono:

- Fattori politici: tasse, politiche monetarie, cambiamenti nei tassi di cambio, ecc. Questo permette di avere informazioni sulla stabilità politica dell'area interessata, sulle politiche fiscali che possono influire sulle operazioni aziendali.
- Fattori economici: come l'inflazione, livelli di disoccupazione, livelli dei tassi di cambio valuta.
- Fattori sociali: reddito medio delle famiglie, abitudini di spesa dei consumatori, struttura demografica. Aiutano nell'identificazione dei potenziali segmenti di mercato e nella valutazione della redditività di un punto di vendita in una specifica zona.
- Fattori tecnologici: sviluppo di nuovi prodotti, cambiamenti nei processi produttivi, innovazioni tecnologiche.

Vediamo nel dettaglio quali possono essere le forze che posso avere un impatto sulle performance delle GDO:

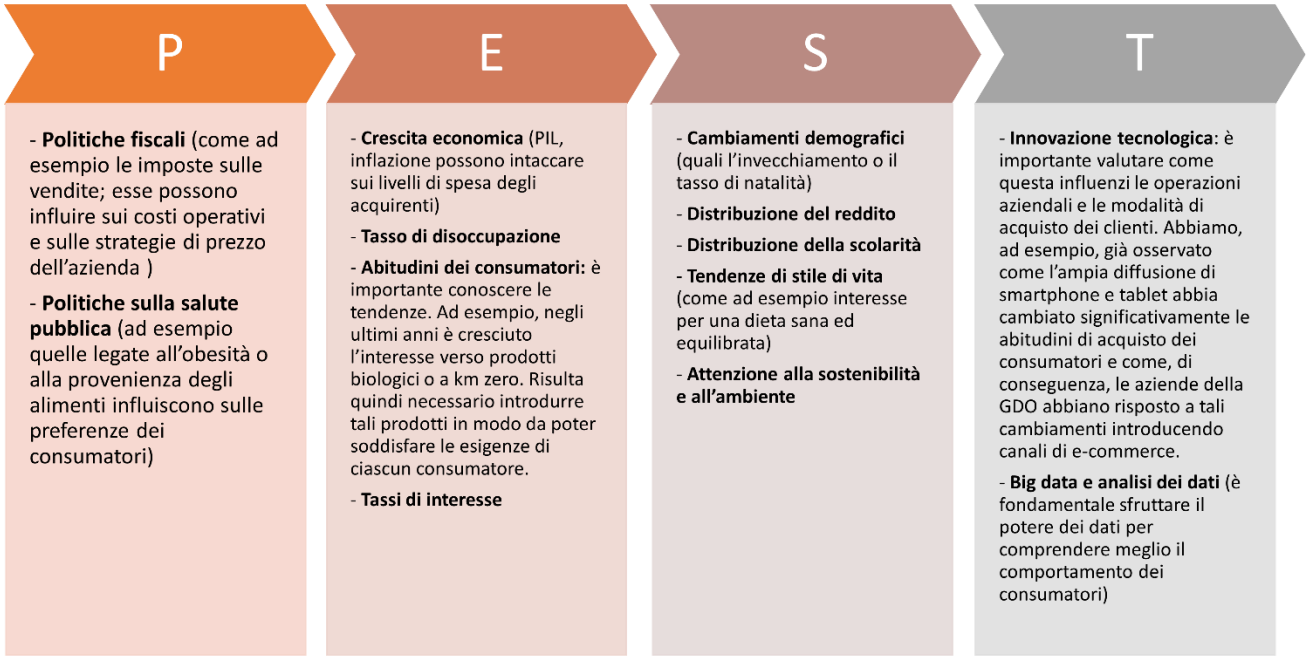

Figura 2: PEST analysis

1.4.3 ANALISI SWOT

Sebbene anche questa analisi non fornisca direttamente informazioni sulla scelta di posizionamento dei nuovi punti vendita, risulta comunque utile per comprendere meglio l'andamento attuale e futuro del proprio business. È uno strumento che aiuta l'azienda a pianificare in modo strategico stando al passo con le tendenze, a identificare i propri punti di forza distintivi, le aree in cui migliorare, nonché le opportunità e minacce che possono influenzare le scelte.

Il termine SWOT sta per:

- S di Strenghts: punti di forza. Rappresentano le iniziative interne che danno risultati positivi.
- W di Weaknesses: punti di debolezza. Tutte quelle iniziative interne che non rendono quanto dovrebbero e che incidono negativamente sulle performance dell'attività.
- O di Opportunities: opportunità. Fattori esterni che possono influenzare positivamente le performance aziendali.
- T di Threats: minacce. Fattori esterni che possono influenzare negativamente l'azienda creando problemi. Generalmente sono fuori dal controllo aziendale ma è possibile comunque ideare un piano di contingenza per ridurne i danni.

I primi due si riferiscono all'ambiente interno, gli altri due a quello esterno. Tuttavia, poiché spesso risulta difficile classificare in punti di forza o debolezza, opportunità o minacce, è consigliabile concentrarsi solamente nelle classificazioni più generiche: fattori interni e fattori esterni. Un limite di questo tipo di analisi è infatti dato dalla difficoltà nel distinguere opportunità e minacce, poiché spesso uno stesso fattore può essere visto in entrambe le prospettive. [4] [5]

Di seguito vengono analizzati alcuni fattori che influenzano la grande distribuzione organizzata:

Figura 3: SWOT analysis

1.4.4 RICERCA DI MERCATO

Un passaggio fondamentale da fare prima di aprire nuove attività/punti di vendita consiste nello svolgere un' indagine di mercato. Con l'espressione "ricerca di mercato" si indica "*lo studio e l'analisi di tutti i fenomeni del mercato rilevanti per qualsiasi tipo di scelta aziendale*" [7]
Lo scopo è quello di riuscire a designare lo scenario nel quale aprire i nuovi punti di vendita, approfondendo contesto economico, comportamenti dei clienti, analizzando aziende competitor già presenti. [6]
Esistono differenti tipologie di ricerca di mercato che vengono racchiuse in due categorie principali: ricerche primarie e secondarie.

RICERCHE PRIMARIE

Il miglior metodo per svolgere tale analisi consiste nel condurre delle indagine qualitative e quantitative, per poi integrare le informazioni ottenute. La ricerca di mercato **quantitativa** riguarda la raccolta di dati statistici, spesso in forma numerica. Serve a compiere stime e misurazioni che,

per essere attendibili, necessitano di un campione di dati molto vasto. Sono esempi di ricerca quantitativa i sondaggi, i questionari, i punteggi delle recensioni. Le ricerche **qualitative** sono relative alla raccolta di dati in forma non numerica. Sono caratterizzate da un approccio più esplorativo, non verranno elaborate delle stime, ma verranno analizzate le componenti emotive, simboliche e cognitive dell'utente finale. Strumenti di ricerca qualitativa sono le interviste singole, i focus group, sentiment analysis . [7]

RICERCHE SECONDARIE

Queste utilizzano dati già raccolti, analizzati e pubblicati. Generalmente sono dati non appartenenti all'azienda, ma provenienti da fonti esterne quali piattaforme istituzionali, fonti governative, organizzazioni di settore. I dati possono includere statistiche demografiche, dati economici, informazioni geografiche, indicatori economici e tanto altro. Queste ricerche possono costituire, in un secondo momento, un supporto per quelle primarie. [8]

Nel presente progetto di tesi vengono condotte numerose ricerche secondarie al fine di individuare la posizione geografica ottimale per l'azienda oggetto dello studio. Queste analisi sono l'esito di una commistione di dati provenienti da diverse piattaforme quali ISTAT (Istituto Nazionale di Statistica), MEF (Ministero dell'Economia e Finanza), AIDA e Statista.

1.4.5 GEO-MARKETING

Per aumentare i benefici che si ottengono tramite l'indagine di mercato è utile affiancargli la tecnica del geo-marketing. Questa metodologia "*permette di fotografare la domanda territoriale attuale e il suo potenziale, valutare la concorrenza, individuare la location più idonea dove aprire la nuova attività e studiare le abitudini d'acquisto dei consumatori di quella specifica area geografica tramite la raccolta di dati che riguardando caratteristiche del target come la residenza, luoghi turistici e città visitate, mezzi di trasporto utilizzati, bar e ristoranti preferiti*" [9] .

Attraverso l'utilizzo di dati geografici, come informazioni demografiche, socioeconomiche, comportamentali e di localizzazione, vengono identificati i target di mercato, le preferenze e i bisogni dei consumatori nelle varie aree geografiche, al fine di rilevare le azioni di marketing specifiche per ciascuna zona. L'output ottenuto da questo processo è un insieme di mappe cognitive e strategiche che correlano informazioni, decisioni e previsioni. [10]

A supporto di questa abbiamo due analisi:

- Analisi spaziale. Utilizzano software GIS (Geographic Information System). Riesce a identificare le aree con una maggiore concentrazione di potenziali clienti e con una minore

presenza di competitor; permette di valutare l'accessibilità ai trasporti e servizi; valuta le caratteristiche sociodemografiche dei vari utenti.

- Analisi del traffico veicolare e pedonale. Viene usata per valutare l'afflusso di persone nelle zone analizzate ottenendo informazioni sulle principali vie di accesso, aree con più pedoni, punti di cogestione e così via, informazioni molto preziose per la scelta della posizione strategica dei nuovi punti vendita

È fondamentale sottolineare che, in un'era così profondamente digitale come la nostra, le imprese spesso si avvalgono di strumenti informatici avanzati per condurre in modo efficiente e accurato tutte queste analisi. L'utilizzo di software e algoritmi vari consente alle aziende di elaborare grandi quantità di dati ottenendo così risultati più dettagliati e specifici. Alcuni strumenti utilizzati dalle GDO sono:

- Software di analisi dei dati
- Data mining e analisi predittiva
- Customer Relationship Management
- Business intelligence

Nel prossimo capitolo, verranno approfonditi tali argomenti, focalizzandoci soprattutto sulla BI.

2. STATO DELL'ARTE TECNICO

2.1 Business intelligence

Oggi, in un mondo altamente dinamico e in continua evoluzione, le aziende si trovano costantemente a confrontarsi con realtà molto diverse dalla loro. Ai fini della loro crescita è essenziale l'analisi e la comprensione del comportamento dei loro competitor. In questo modo, le aziende riescono ad acquisire una visione più completa del contesto in cui operano e conseguentemente possono prendere decisioni più ponderate e strategiche.

L'utilizzo di strumenti di Business Intelligence permette alle imprese di implementare la loro conoscenza in merito alle proprie dinamiche interne e, contestualmente, maturare una migliore cognizione riguardo il funzionamento delle altre aziende operanti nello stesso mercato. Ciò si tramuta in un vero e proprio vantaggio competitivo.

L'espressione "Business Intelligence" (BI) indica un insieme di processi e tecnologie aziendali che si basano sull'analisi dei dati. Attraverso operazioni di raccolta, elaborazione e valutazione di questi dati, vengono fornite informazioni utili al servizio del management strategico, in modo da poter permettere alle organizzazioni di prendere decisioni basate sui dati, di migliorare le prestazioni, indentificare eventuali problemi e individuare le tendenze di mercato. (Rezzani, 2012)

Lo sviluppo del sistema di BI si articola in quattro strati differenti: (Noce & D'Ercole, 2000)

1. Strato delle sorgenti, in cui si identificano le varie fonti. La fonte di provenienza dei dati è solitamente uno o più database tradizionali e gestionali eterogenei (come ERP, MES, PLN, CRM), dai quali vengono raccolti - in modo automatizzato- direttamente dall'ambiente operazionale. Di recente, tuttavia, si tende ad attingere anche a sorgenti destrutturate (dati che provengono da sorgenti alternative, quali ad esempio attività social, siti web, e così via). L'utilizzo di queste sottolinea l'importanza e la necessità di integrare i dati in unico ambiente allineato e uniforme.
2. Strato di alimentazione, noto anche come ETL (estrazione, trasformazione e caricamento). In questa fase i dati vengono estratti dalle sorgenti e successivamente convertiti dal formato operazionale sorgente a quello del DWH.
3. Strato del Datawarehouse, i dati standardizzati verranno integrati in un unico ambiente di archiviazione (Datawarehouse). È il livello che si occupa della gestione del dato per l'intero ciclo di vita della base di dati.
4. Strato di analisi, fase in cui vengono effettuate le analisi di dati.

I requisiti che una BI deve avere sono:

- Velocità, i sistemi devono essere in grado di leggere ed elaborare i dati rapidamente, anche quando questi sono presenti in grandi quantità;

- Facilità d'uso, deve utilizzare un linguaggio di facile comprensione e immediato nell'interpretazione;
- Integrazione, i dati raccolti provengono da molteplici fonti e possono essere di diversa tipologia (ad esempio numerici, di testo, immagini e così via). È quindi necessario saper trattare e integrare questi tipi differenti di dati in modo da poter creare un insieme completo e coerente di informazioni;
- Storicizzazione, deve mantenere la storia dei cambiamenti subiti da certi attributi selezionati, per permettere analisi storiche contestualizzate (D'Ovidio & Villari, 2009)

2.2 Big Data

I Big Data si riferiscono a un dataset la cui dimensione va al di là della capacità di un database normale di catturare, memorizzare, gestire e analizzare i dati (Manyika & M, 2011). Questo termine viene usato per descrivere un insieme di dati che provengono da diverse fonti sia interne che esterne all'azienda. Oggi, come detto precedentemente, le aziende si ritrovano a generare continuamente dei dati, che per essere utilizzati devono prima essere storicizzati mediante l'utilizzo di strutture quali il Data Warehouse.

Ciò che caratterizza tutti i Big Data sono le cosiddette 5V: volume, velocità, varietà, veridicità e valore. [11]

- Volume, si riferisce alla grande quantità di dati generati attraverso numerosi canali
- Velocità, ovvero la rapidità con la quale i dati vengono acquisiti e utilizzati grazie a transazioni sempre più frequenti e veloci
- Veridicità, riguarda la qualità dei dati e il loro livello di sicurezza
- Valore, l'obiettivo principale è quello di creare valore dai dati attraverso l'estrazione di informazioni utili e insight
- Varietà, è legata alle tipologie di dati differenti che possono essere distinti in:
  - Strutturati: dati conservati in Database relazionali, organizzati secondo schemi e tabelle rigide
  - Non strutturati: dati conservati senza alcuno schema (come, ad esempio, forme libere di testo stralci di e-mail, articoli, audio senza tag e così via)
  - Semi-strutturati: dati che presentano caratteristiche di entrambe le tipologie succitate. (Raghupathi & V., 2014)

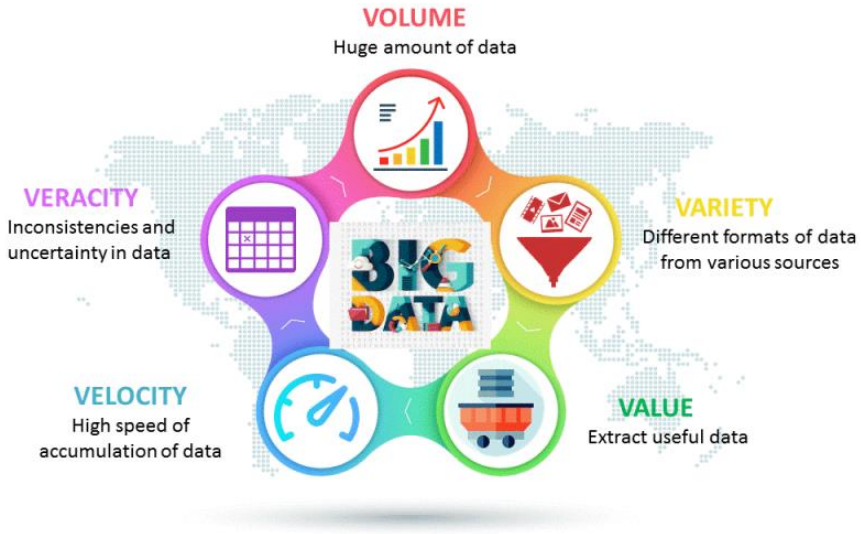

Figura 4: Le 5V dei Big Data

2.2.1 Benefici dei Big Data

Lato analytics troviamo diversi vantaggi:

- Supporto nel processo di decision making e previsioni più precise: con l'utilizzo di analytics sofisticati su interi database è possibile automatizzare e migliorare i processi decisionali tramite le predizioni dei KPI, ridurre i rischi e scoprire insight. (Manyika & M, 2011)
- Creare trasparenza: grazie alla facilità e tempestività con cui si può accedere ai Big Data, è possibile ottenere una vasta quantità di informazioni che possono essere condivise in modo semplice tra le diverse unità organizzative di una impresa. (Bernice, 2013)
- Profiling dei consumatori: la disponibilità di dati quasi in real time, ottenuti tramite dispostivi come gli smartphone, consente di ottenere dettagli sul processo decisionale dei consumatori durante l'acquisto. Grazie all'analisi di queste informazioni si è in grado di identificare i modelli comportamentali dei consumatori e riconoscere le loro esigenze.
- Aumento dell'efficacia e dell'efficienza aziendale: lo sfruttamento dei big data può aiutare le imprese a realizzare più output diminuendo il numero di input, migliorando anche il livello di qualità dell'output stesso. In questo modo viene incrementata non solo la produttività ma anche la profittabilità delle aziende. (Manyika & M, 2011) (McAfee A. 2012)

Numerosi sono i benefici che le imprese possono ottenere grazie al corretto utilizzo dei Big Data, tra i quali:

• Miglioramento delle prestazioni: la creazione e memorizzazione dei dati transazionali in forma digitale fornisce alle aziende dati più accurati e precisi in merito alle diverse performance aziendali, perfino in tempo reale o quasi. (Manyika & M, 2011)

• Personalizzazione delle azioni: la gestione dei Big Data consente la formazione di segmenti di clientela, definiti cluster. Questi risultano utili per personalizzare prodotti e/o servizi e offrire promozioni e pubblicità specifiche alle esigenze di ciascun cluster. (Manyika & M, 2011)

Tuttavia, nonostante questi benefici e opportunità, alcune aziende si mostrano ancora scettiche nell'affidarsi ai Big Data. Ciò può derivare, ad esempio, dalla difficoltà nel comprendere i nuovi strumenti o nelle ingenti spese (in termini sia di tecnologie da implementare che di nuove figure da dover assumere) che queste iniziative richiedono. (Court, 2015)

2.2.2 Tecniche per l'analisi dei Big Data

Esistono varie tecniche per aggregare, manipolare, gestire e analizzare i dati. Di seguito vengono elencate le principali:

- A/B testing: viene confrontato un gruppo di controllo con gruppi test al fine di determinare quali modifiche e azioni possano migliorare una data variabile obiettivo;
- Crowdsourcing: tecnica usata per raccogliere dati da un vasto gruppo di individui, solitamente tramite piattaforme online;
- Data integration: metodologie volte a integrare e analizzare dati provenienti da fonti diverse al fine di ottenere insight più affidabili ed efficienti rispetto a quelli che si ottengono esaminando una singola fonte;
- Data mining: disciplina che utilizza una combinazione di tecniche di classificazione, cluster analysis, regole associative e regressione per scoprire modelli, relazioni e informazioni significative all'interno di grandi dataset.
- Machine Learning: sottocampo della computer science riguardante la progettazione e creazione di algoritmi capaci di consentire ai sistemi informatici di acquisire conoscenza dai dati; in questo modo il computer sarà in grado di identificare modelli complessi all'interno dei dati e di prendere decisioni o fare previsioni autonomamente.
- Natural Language processing: campo della AI che consente alle macchine di comprendere, interpretare e generare il linguaggio usato dagli esseri umani, con lo scopo di permettergli di analizzare ed elaborare testi o discorsi in modo simile a come verrebbe fatto da un umano;
- Regressione: tecniche che individuano come il valore di una variabile dipendente cambia al variare di una o più variabili indipendenti;

- Ottimizzazione: tecniche volte a migliorare le performance relative a diversi aspetti, quali ad esempio: costi, velocità o affidabilità, attraverso la riprogettazione di sistemi e processi complessi.
- Sentiment analysis: volta a valutare l'opinione, l'atteggiamento o il sentimento espresso in un testo (come una recensione, un post sui social media, un feedback). L'obiettivo è quello di identificare la tipologia del sentimento (positivo, negativo, neutro) nei confronti di un determinato argomento;
- Statistica: disciplina usata per raccogliere, organizzare e interpretare i dati al fine di trarre conclusioni sulle relazioni tra le varie variabili, così da distinguere quelle verificate per puro caso da quelle invece che hanno una base causale (statisticamente significative);
- Data visualization: processi di rappresentazione grafica dei dati e delle informazioni. Vengono create immagini, diagrammi, grafici, mappe al fine di facilitare la comprensione e l'interpretazione delle analisi fatte;
- Modelli predittivi: usano modelli matematici per prevedere la probabilità di un risultato;

Nell'ambito dei Big Data sono state introdotte molteplici tecnologie. Tra queste, rivestono un ruolo particolarmente importante i Data Warehouse, i Data Lake e i Data Lakehouse:

- Il Data Warehouse è un sistema di archiviazione dei dati che, prima dell'inserimento vengono sottoposti ad un processo di ETL (estrazione da varie fonti aziendali, trasformazione e infine caricamento nel Data Warehouse). La struttura dei dati viene definita in anticipo, facilitando in questo modo l'analisi dei dati storici.
- Il Data Lake, invece, archivia e memorizza i dati nel formato originale, senza che subiscono alcuna trasformazione. È progettato per accogliere una vasta gamma di dati. La sua struttura offre un approccio più flessibile, consentendo un accesso rapido e diretto alle varie informazioni.
- Il Data Lakehouse è invece la combinazione di entrambi i concetti, combinando in questo modo la flessibilità e l'agilità del Data Lake con la struttura e le prestazioni del Data Warehouse.

2.3 Data Lake

L'obiettivo di tale struttura è memorizzare una grande quantità di dati nel loro formato nativo. Esistono diversi stadi di sviluppo di un Data Lake. La fase iniziale è chiamata "*data puddle*"

(pozzanghera di dati). In questa fase i dati vengono raccolti senza una struttura o organizzazione definita. Solitamente sono raccolte di modeste dimensioni di proprietà di un unico team.
Successivamente, si progredisce al "*data pond*" (stagno di dati). In questa fase, i dati iniziano ad essere organizzati e categorizzati in base a determinati criteri, rappresentando una raccolta di data puddle.
Abbiamo poi il *data lake* (lago di dati) vero e proprio. In questa fase i dati sono raccolti in un'unica piattaforma e vengono ben documentati. A differenza del data pond esso supporta il self-service (ovvero i diversi utenti aziendali sono in grado di trovare e utilizzare i dati senza chiedere aiuto al reparto IT) e, inoltre, contiene informazioni che non necessariamente servono in quel momento ma potrebbero servire per progetti futuri.
Nel momento in cui il data lake si espande notevolmente in termini di dimensioni si crea il cosiddetto *data ocean* (oceano di dati). Esso contiene una grande quantità di dati provenienti da origini differenti , permettendo così di ottenere una visione completa e dettagliata dei dati aziendali.
A volte, però, i data lake presentano una struttura disorganizzata , creando i *data swamp* (palude di dati). I dati presenti in un data swamp spesso sono obsoleti o privi di struttura, rendendo in questo modo difficile l'elaborazione .

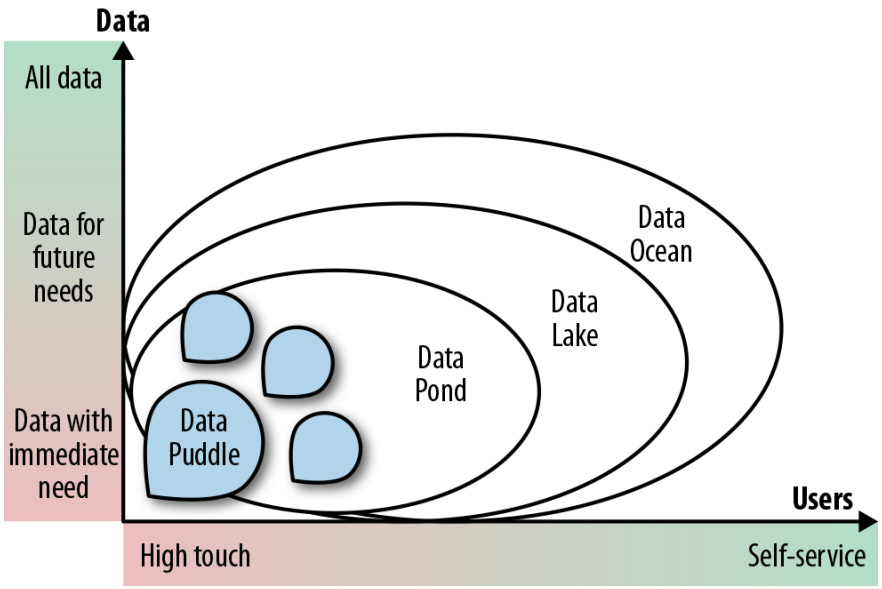

Figura 5: Le 4 fasi della maturità

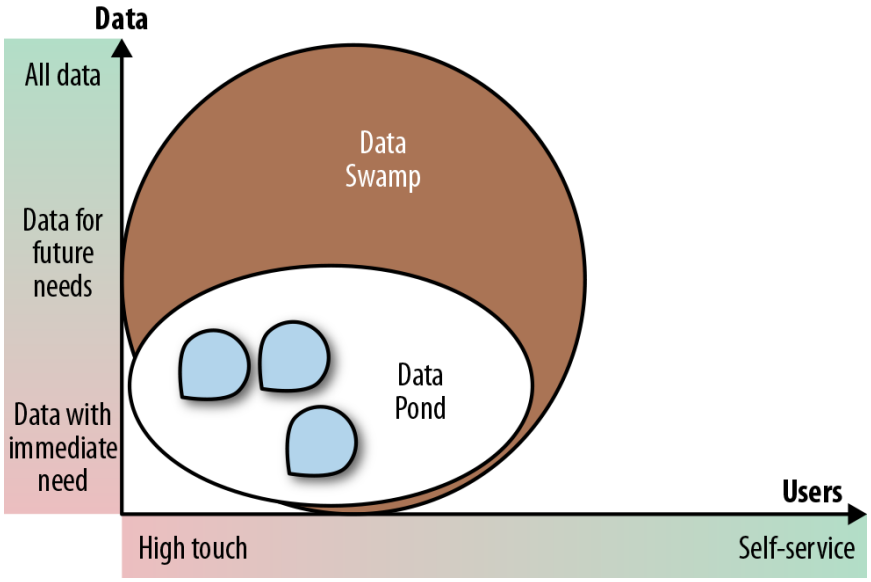

Figura 6: Data swamp

Possiamo quindi affermare che adottare i Data Lake comporta diversi benefici, tra cui:

- Maggiore flessibilità: conservando i dati nel loro formato originario è possibile archiviare una vasta gamma di tipi di dati eterogenei senza dover fare alcuna elaborazione o trasformazione preliminare. Questa caratteristica favorisce una gestione più flessibile dei dati e consente di adattarsi più facilmente alle diverse analisi;
- Ampia scalabilità: riescono a supportare un volume sempre maggiore di dati;
- Facilità nell'esplorazione dei dati: essendo questi conservati in forma grezza i Data lake permettono una esplorazione senza vincoli rigidi. In questo modo riescono ad eseguire le varie analisi senza dover definire uno schema predefinito;
- Riduzione dei costi: non dovendo adottare alcuna trasformazione vengono evitati i costosi processi di trasformazione.[12]

2.4 Datawarehouse

Un Data Warehouse (DWH) è un tipo di sistema di data management progettato per abilitare e supportare le attività di business intelligence (BI), in particolare gli analytics. I DWH consentono di raccogliere dati integrati, consistenti e certificati, riguardanti tutti i processi di business dell'azienda e provenienti dalle fonti operazionali. I dati verranno successivamente trasformati attraverso procedure di ETL e sottoposti a controlli di data Quality. Questo ultimo passaggio è di fondamentale importanza, poiché se i dati dovessero risultare "sporchi" potrebbero provocare un peggioramento delle performance aziendali, con conseguente aumento di costi e perdite di opportunità per l'azienda.

Il DWH può essere centralizzato o decentralizzato. Si definisce centralizzato quando è presente un unico database fisico contenente tutti i dati utili al supporto decisionale. In questo modo si può ottenere un maggiore controllo sulla qualità dei dati e l'uniformità delle informazioni; tuttavia, potrebbe essere costoso tanto da implementare quanto da gestire. Si parla, invece, di DWH decentralizzato quando i dati vengono distribuiti su più server o nodi all'interno di una rete. Questo rende più semplice la condivisione delle informazioni tra le varie funzioni aziendali, ma potrebbe portare a problemi di coerenza dei dati.
A causa dell'ingente numero dei dati presenti, spesso per alleggerire il carico di lavoro si decide di dividerli in differenti aree tematiche, i Data Mart. Un Data Mart, dal momento che si configura come parte di un sottoinsieme di un Data Warehouse, svolge le stesse funzioni di quest'ultimo ma è limitato ad ambiti più specifici e si riferisce a determinate linee di business o a un determinato reparto. Questo li rende più semplici da definire e consentono di scoprire rapidamente degli insight più mirati, rispetto a quando si lavora con dataset più ampi. Il contro è che tuttavia possono introdurre incoerenza, poiché può risultare complicato la gestione e il controllo dei dati presenti in numerosi data Mart. [13]

Le caratteristiche che deve avere un DWH sono:

- Orientato al soggetto: i dati devono essere organizzati per soggetti rilevanti (prodotti, clienti, fornitori, date, ecc.) in modo da offrire tutte le informazioni inerenti a una determinata area
- Integrato: deve essere in grado di integrarsi al meglio con i differenti standard utilizzati nelle varie applicazioni. I dati devono infatti essere ricodificati al fine di risultare omogenei dal punto di vista semantico e devono utilizzare le stesse unità di misura, in modo da eliminare ogni tipo di incompatibilità
- Variabile nel tempo: i dati presenti nel DWH presentano un orizzonte temprale molto ampio e possono essere riutilizzati in diversi istanti temporali
- Non volatile: le informazioni nel tempo devono mantenere la loro integrità, devono essere conservati permanentemente (Inmon, 1992)

2.4.1 Progettazione di un Datawarehouse

Il primo passo da fare per progettare un DWH è la definizione degli specifici requisiti aziendali e la progettazione concettuale, che comprende sia l'aspetto fisico (come archiviare e recuperare gli oggetti) che logico (riguardante le relazioni tra gli oggetti). Nella fase di progettazione è fondamentale considerare le esigenze degli utenti finali. Tuttavia, questi non sempre sanno cosa vogliono o di cosa hanno bisogno. È quindi importante che il processo di pianificazione e

progettazione includa una esplorazione adeguata in modo da poter anticipare eventuali esigenze future degli utenti. [14]

Dopo aver stabilito gli obiettivi e i bisogni degli utenti si può procedere alla fase di progettazione logica. Questa comprende la scelta dello schema di progettazione più adatto. Gli schemi sono una delle modalità usate per organizzare i dati all'interno del data Warehouse. Le due strutture principali sono lo star schema e lo snowflake. [15]

2.4.2 Star schema

La struttura prevede una tabella centrale, la tabella dei fatti, che può essere unita a diverse tabelle de-normalizzate, dette "delle dimensioni". Il fatto è un evento di interesse per l'azienda (vendite, spedizioni, ...) e sono solitamente valori numerici e quantitativi, mentre le dimensioni contengono le informazioni descrittive degli elementi della tabella dei fatti (prodotto, punto vendita, data, …).
Il maggior beneficio di questa struttura è la facilità con cui si possono ricercare i valori desiderati. Questo è possibile grazie alla presenza di query scritte con pochi e semplici inner join tra tabella dei fatti e quelle delle dimensioni. Lo svantaggio è invece dato dalla possibile ridondanza dei dati all'interno delle dimensioni, che causa un aumento dello spazio di archiviazione necessario. [16]

2.4.3 Snowflake schema

La struttura dello snowflake può essere intesa come un'estensione dello star schema. La differenza risiede nella struttura, infatti nello snowflake è presente sempre una tabella dei fatti al centro collegata con le varie tabelle delle dimensioni, ma queste ultime presentano anche delle ramificazioni con altre tabelle normalizzate, spesso su più livelli (la cui rappresentazione grafica ricorda appunto proprio i fiocchi di neve). I benefici di questa struttura sono:

- Assenza di ridondanza dei dati
- Aggiornamento dei dati più semplice
- Minore spazio di occupazione per la conservazione dei dati
- Maggiore velocità di accesso alle informazioni

Il contro invece è dato dalla complessità delle query, causato dall'utilizzo di più join rispetto allo star schema. [17]

È però importante sottolineare che, a prescindere dai pro e contro di ciascuno degli schemi analizzati, la scelta di uno piuttosto che l'altro dipende principalmente dalle specifiche esigenze dell'azienda e dalle caratteristiche dei dati da dover memorizzare.

2.5 ETL

Come già visto, tra i livelli di Business Intelligence troviamo il livello di alimentazione. Questo processo prevede tre fasi principali: estrazione, trasformazione e caricamento dei dati (da questo l'acronimo inglese ETL: extract, load, transform). È una procedura di fondamentale importanza perché non tutte le informazioni che vengono raccolte possono essere immediatamente memorizzate, alcune hanno bisogno di modifiche o di essere riportate al formato di compatibilità.
Il ruolo degli strumenti di ETL consiste nell'alimentazione di una sorgente dati singola o multipla, dettagliata, esauriente e di alta qualità che a sua volta dovrà alimentare il DWH. Le operazioni svolte da questi processi prendono il nome di riconciliazione. [18]
Tra i vari casi d'uso dei processi ETL troviamo: [19]

- Il trasferimento dei dati da una applicazione a un'altra
- La replica dei dati per eseguire un backup o l'analisi della ridondanza
- Acquisizione dei dati provenienti da varie fonti e integrazione in un DWH per il loro utilizzo nell'ambito di BI
- Migrazione delle applicazioni locali in ambienti Cloud, Cloud ibride o multi-cloud

Come precedentemente detto le fasi che compongono questo livello sono:

- Estrazione
- Pulitura
- Trasformazione
- Caricamento

Bisogna dire che sta diventando molto comune la pratica ELT (extract, load, transfrom). I due processi differiscono su due punti principali: quando e dove avviene la trasformazione. Entrambi richiedono delle aree di staging (aree di transito che sperano le fasi di estrazione e trasformazione da quelle di caricamento), ma mentre negli ETL queste aree si trovano tra il sistema di origine e il sistema di destinazione (il data warehouse), e quindi le trasformazioni vengono fatte prima del caricamento dei dati, negli ELT l'area di staging si trova proprio nel DWH, permettendo di eseguire le trasformazioni all'interno di esso. [20]

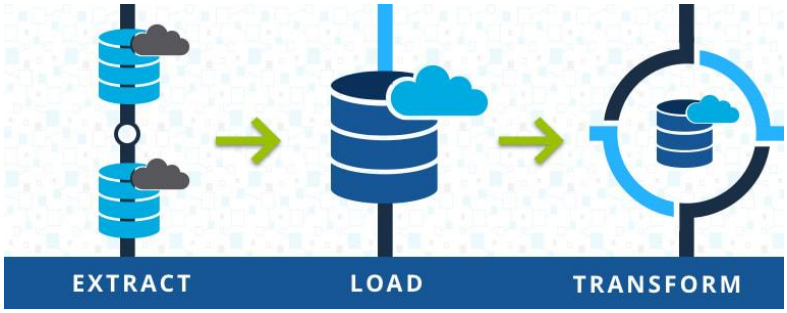

Figura 7: ELT

Figura 8: ETL

2.5.1 Estrazione

L'estrazione dei dati è di due tipi, statica (IL) e incrementale (CDC):

- Inital Load (IL): Viene effettuata all'inizio, quando il DWH viene popolato per la prima volta e consiste concettualmente in una fotografia dei dati operazionali
- Change Data Capture (CDC): Serve per aggiornare un periodo del DWH, infatti si utilizza per catturare esclusivamente le modifiche apportate ai dati dopo l'ultima estrazione. Questo consente di avere un tempo di estrazione minore nonché un DWH aggiornato in modo efficiente.

2.5.2 Pulitura e trasformazione

In generale l'operazione di pulitura è associata alla correzione dei valori dei dati mentre la trasformazione si occupa del loro formato.

Nel dettaglio, la pulitura si incarica di migliorare la qualità dei dati provenienti dalle varie sorgenti, in modo da averne un insieme coerente e affidabile, così da poterli usare a supporto delle decisioni aziendali. Ha quindi lo scopo di correggere eventuali problemi, quali: dati duplicati, dati mancanti, valori errati o inconsistenti e così via.

La trasformazione, invece, si occupa della conversione dei dati provenienti dalle diverse fonti sorgente al formato del datawarehouse. Questa avviene tramite l'utilizzo di una serie di regole definite a livello aziendale che servono a garantire l'accessibilità e qualità dei dati. Queste regole includono:

- Standardizzazione, per definire quali dati verranno presi in considerazione, come saranno formattati e memorizzati. Questo processo è cruciale per garantire affidabilità e coerenza dei dati
- Deduplicazione, per eliminare eventuali ridondanze
- Verifica, per mettere a confronto informazioni simili tra di loro e segnalare eventuali anomalie
- Ordinamento, per ottimizzare l'efficienza del DWH i dati vengono raggruppati e ordinati in categorie di elementi [19]

2.5.3 Caricamento

Ci sono due differenti modalità di caricamento dei dati nel DWH:

- Refresh: tecnica usata principalmente durante la fase iniziale di popolamento, consiste nella reiscrizione integrale dei dati con la conseguente sostituzione completa di quelli precedenti.
- Update: è usata invece insieme all'estrazione incrementale per l'aggiornamento periodico. Infatti, vengono aggiunti nel DWH solo i cambiamenti avvenuti nei dati

Per accelerare il processo di ETL e ridurre il tempo di caricamento dei dati è possibile adottare la tecnica di parallelizzazione. Per fare ciò ci sono due approcci differenti:

- Passi di carico multipli: il flusso di lavoro viene suddiviso in più lavori indipendenti che possono essere eseguiti in contemporanea. Ogni attività deve essere in grado di gestire eventuali errori in modo autonomo
- Pipeline: il database identifica automaticamente i diversi compiti che può eseguire in parallelo.

3. TRADITIONAL ETL PER LA CREAZIONE DEL DATA MART

L'integrazione dei dati, come già precedentemente detto, è il processo volto a combinare dati provenienti da diverse origini al fine di offrire agli utenti una unica visualizzazione. [21]
Inizialmente, le aziende hanno risposto alla necessità di dati integrati sviluppando al loro interno software ad hoc per eseguire le fasi di estrazione, trasformazione e caricamento dei dati in un ambiente unico e integrato.

In seguito, con l'aumento dell'importanza dei dati e l'aumento della complessità delle architetture nasce la necessità di avere strumenti più efficienti. Nascono così i primi software proprietari sviluppati da aziende specializzate nell'integrazione dei dati. Tuttavia, nonostante i progressi ancora oggi molte aziende utilizzano soluzioni ETL personalizzate.
Negli ultimi anni, inoltre, la diminuzione dei budget assegnati allo sviluppo dell'Information Technology, causata dalla recente crisi economica, ha determinato una maggiore richiesta di soluzioni open source.
Il mercato della data Integration si può quindi dire caratterizzato da tre tipologie differenti di prodotti:

- Software personalizzati: rappresentano tutti i prodotti sviluppati internamente alle aziende per rispondere all'incrementale bisogno di integrare i dati. Questo approccio sta però diventando sempre meno conveniente.
- Software proprietari: vengono sviluppati da aziende specializzate nel campo. Il numero degli applicativi oggi è molto elevato. Si hanno prodotti che coprono quasi tutte le necessità aziendali e prodotti più specifici da usare in determinati contesti di business o per particolari problematiche. Tra i più diffusi troviamo: IBM InfoSphere DataStage, SAP Data Service, Microsoft SQL Server Integration Services (SSIS), Oracle Data Integrator (ODI).
- Software open source: uno degli svantaggi dell'utilizzo dei software proprietari è rappresentato dagli elevati costi di implementazione. Nascono così i prodotti open source. Questi supportano una discreta quantità di funzioni ma con un costo decisamente minore rispetto a quelli proprietari. Ne sono l'esempio Talend Open Studio, Pentaho Data Integration (Kettle). (Bregata, 2019)

Per questo progetto di tesi si è deciso di utilizzare IBM DataStage.

3.1 IBM INFOSPHERE DATASTAGE

Tra i più diffusi strumenti di integrazione dei dati troviamo quello sviluppato da IBM: DataStage. DataStage è riconosciuto come leader nel settore di integrazione dei dati, come confermato dalla valutazione del quadrante di Gartner (guarda figura 9). Quest'ultimo è dato da una serie di report che analizzano e classificano i principali fornitori di soluzioni di data integration.

DataStage è uno strumento che aiuta a progettare, sviluppare ed eseguire le attività di spostamento e trasformazione dei dati, supportando sia i modelli ETL che quelli ELT. Con la sua interfaccia grafica (GUI) intuitiva semplifica la creazione di pipeline all'utente e agevola l'accesso, la trasformazione e lo spostamento dei dati. [22]

Source: Gartner (August 2022)

Figura 9: Magic Quadrant per i tools di Data Integration

In dettaglio, DataStage offre diverse funzionalità significative:

- È uno strumento potente che facilita l'implementazione dei processi di ETL.

- È ideale per progetti di integrazione dei dati, come ad esempio la creazione di un data Warehouse o un Data Mart.
- Consente l'importazione, l'esportazione, la creazione e gestione dei metadati utilizzati a supporto dei job.
- Permette la pianificazione, l'esecuzione, il monitoraggio dei processi, garantendo in questo modo una gestione efficiente e semplificata dei processi ETL.
- Viene utilizzato anche per amministrare gli ambienti di sviluppo ed esecuzione, inclusa la gestione delle autorizzazioni degli utenti, la configurazione delle connessioni ai database di origine e destinazione, il monitoraggio delle varie risorse.
- Offre la possibilità di creare lavori batch (lavori eseguiti automaticamente senza la necessità dell'intervento dell'utente)

L'architettura è basata su un modello Client-Server, dove il *client* rappresenta l'interfaccia grafica mentre il *server* è il motore che esegue i flussi dei dati e gestisce la trasformazione, il caricamento e lo spostamento dei dati.

Sul lato client troviamo i seguenti componenti:

- DataStage Administrator: gestisce i progetti di DataStage e svolge la manutenzione del server. È responsabile della configurazione e dell'amministrazione del sistema.
- Designer: serve per la creazione dei job
- Director: esegue e monitora i vari processi.
- Information server manager: fornisce una interfaccia utente che permette agli amministratori di gestire, configurare e monitorare i vari componenti di DataStage

I vantaggi legati all'utilizzo di questo tool sono:

- **Riduzione dei costi** di spostamento dei dati.
- **Consegna dati attendibili**, grazie all'utilizzo di funzioni di governance di IBM Cloud Pak for Data.
- **Ampia connettività** con una vasta gamma di connettori per integrare i dati provenienti da database relazionali, sistemi mainframe, servizi web e così via.
- **Potenti funzionalità di trasformazione**, grazie all'interfaccia grafica intuitiva che definisce facilmente le varie trasformazioni. Inoltre, sono presenti vari set predefiniti quali filtri, aggregazioni, join che permettono una semplice manipolazione dei dati.
- **Scalabilità e prestazioni elevate** in grado di gestire grandi volumi di dati. Inoltre, l'utilizzo dell'elaborazione parallela garantisce un'esecuzione dei carichi di lavoro più veloce.
- **Gestione avanzata dei metadati,** semplificando la documentazione, la comprensione e la manutenzione dei processi di integrazione.

- **Funzionalità di monitoraggio e controllo**, consentendo di tracciare lo stato e le prestazioni delle attività di integrazione in tempo reale. Questo facilita nella risoluzione dei problemi e nell'ottimizzazione delle prestazioni.
- **Integrazione con l'ecosistema IBM** ,DataStage riesce ad integrarsi senza problemi con altri prodotti IBM consentendo così una integrazione più completa e coerente all'interno di tutto l'ecosistema di IBM.

Nel mercato esistono più versioni che rispondono alle diverse esigenze delle aziende. Per le piccole-medie imprese che manipolano un numero limitato di dati si ha ad esempio il Workgroup Edition; per le realtà medio-grandi che invece elaborano un numero elevato di dati in un tempo limitato si ha l'Enterprise Edition. Negli ultimi anni, c'è stato un crescente interesse anche nell'opzione di utilizzare DataStage in ambiente cloud, come il Cloud Pak for Data. (IBM, 2010) [23]

3.2 Creazione del Data Mart

Un Data Mart, come già menzionato nello stato dell'arte, è un sottoinsieme di un DWH che si focalizza su specifiche aree di business. La creazione di un Data Mart comporta numerosi vantaggi, tra cui:

- Efficienza in termini di costo
- Facilità nella reperibilità dei dati: l'utente riesce a recuperare con meno sforzi i dati in un data Mart rispetto a un DWH.
- Rapidità nell'accedere agli insight: identificare dati di valore in un lasso di tempo minore consente di prendere decisioni strategiche più velocemente e migliora la produttività complessiva.
- Manutenzione dei dati più semplice: una quantità inferiore di informazioni aziendali rispetto a un DWH comporta una manutenzione più facile
- Minori tempi di implementazioni

[24]

Esistono due approcci differenti di implementazione:

- TOP-DOWN: prevede prima la creazione di un DWH contenente tutti i dati aziendali. Successivamente vengono aggregati e organizzati così da creare i vari Data Mart specifici per ogni area di business. Si ottiene in questo modo una visione aziendale completa, offrendo anche la possibilità di analizzare i dati in prospettive differenti.
- BOTTON-UP: vengono creati i Data Mart corrispondenti a ciascuna area di business interessata. Successivamente i Data Mart possono essere aggregati così da creare il DWH. È

utile utilizzare questo approccio quando si desidera ottenere velocemente le informazioni su determinate aree di business.

La scelta del tipo di approccio da utilizzare dipende dalle esigenze specifiche dell'organizzazione e dalla complessità dei dati a disposizione. [25]

L'obiettivo di questo elaborato è identificare le aree geografiche per l'apertura di nuovi punti di vendita. A tale scopo, è stato progettato un Data Mart che consente di combinare in modo efficiente e coerente le informazioni aziendali con dati provenienti da fonti aperte. L'approccio usato per l'implementazione è quello di tipo *top-down*.

Come primo passo, sono stati selezionati i dati interni all'azienda, come ad esempio il fatturato, le vendite, le informazioni sui punti vendita già esistenti. Questi dati costituiscono la base per comprendere le prestazioni attuali dell'azienda.

Successivamente, sono stati acquisiti i dati provenienti da fonti open, quali ad esempio ISTAT, Ministero dell'Economia e delle Finanze, Statista e altre fonti rilevanti per lo studio. Questi includeranno tutte le informazioni utili sulle zone di interesse, come informazioni demografiche, dati socioeconomici, dati di mercato.

Una volta che i dati, sia interni che esterni, sono stati integrati nel Data Mart è possibile procedere con la cross analysis. Questa analisi identifica le relazioni e correlazioni significative tra le diverse variabili, permettendo di comprendere come una variabile possa influenzare l'altra o come le variabili si influenzano reciprocamente.

Attraverso questa analisi, saranno quindi fornite raccomandazioni solide e basate sui dati utili per l'espansione strategica aziendale, al fine di identificare le posizioni geografiche che presentano un potenziale di mercato promettente per l'apertura di nuovi punti vendita.

Durante tutta l'analisi risulta fondamentale garantire integrità e coerenza dei dati, per questo tutte le transazioni dovranno seguire rigorosamente il sistema ACID. ACID deriva dall'acronimo inglese Atomicity, Consistency, Isolation, e Durability:

- Atomicità: la transazione deve essere eseguita in modo atomico, come se fosse un'unità indivisibile. Ovvero, al suo termine, se tutti i passaggi vengono eseguiti con successo, i suoi effetti devono essere resi permanenti e visibili nel database; invece, in caso di fallimento nessun effetto deve poter essere mostrato e i dati devono essere ripristinati allo stato precedente, come se la transazione non fosse mai avvenuta.
- Consistenza: quando si conclude una transazione devono essere soddisfatti i vincoli di integrità, cioè il database deve essere in uno stato coerente sia prima che dopo la transazione.

- Isolamento: ogni transazione deve necessariamente essere eseguita in modo indipendente dalle altre al fine di non interferire con il risultato le altre transazioni in corso.
- Durabilità: i risultati ottenuti dalle transazioni devono essere permanenti e persistenti nel sistema. [26]

Il processo ETL eseguito è diviso in tre livelli:

- L0: detto anche *staging area*. Qui vengono estratti i dati da varie tipologie di file senza alcuna trasformazione.
- L1: *operational data store.* Vengono eseguite le principali trasformazioni e le operazioni di data Quality.
- L2: *publication area.* Livello di pubblicazione dei dati conformi all'analisi di business; rappresenta infatti l'area da cui legge il front-end.
- L2: cross analysis

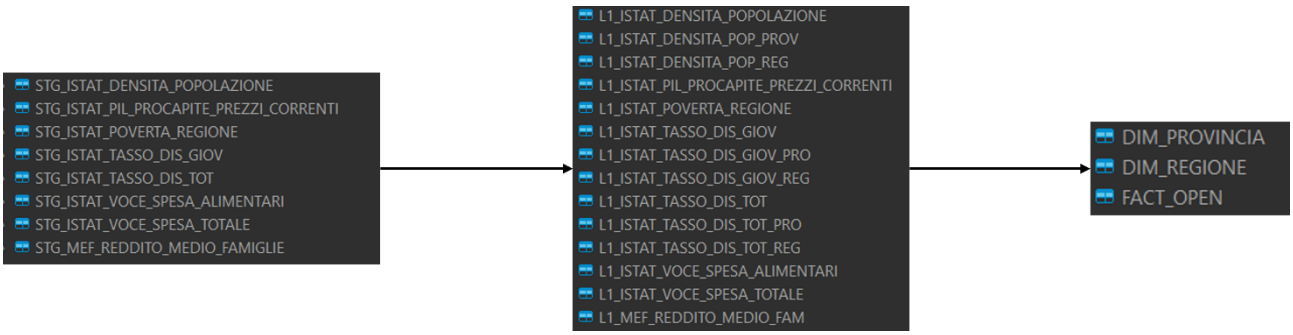

Figura 10: Processo ETL

3.2.1 Delta dei dati

Come sappiamo, l'alimentazione del DWH inizia con il processo *Initial Load* (IL). Come si intuisce dal termine stesso, questo prevede il popolamento iniziale del DWH in cui tutti i dati vengono estratti dalle fonti sorgenti e trasferiti per la prima volta nel data Warehouse. Dopo questa prima fase si passa al *Delta Load,* caricamento incrementale, in cui vengono identificati e catturati esclusivamente i dati che hanno subito cambiamenti o dati aggiunti dopo il processo iniziale.

Diversi sono gli approcci da poter eseguire:

- CDC (change data capture): meccanismo che permette di intercettare nelle tabelle il delta dei dati rispetto alla precedente estrazione. È usato principalmente per la replica di tabelle molto grandi.
- MINUS: vengono estratte autonomamente tutti i dati al fine di fare una minus dei dati che si hanno con quelli nuovi

- FULL: viene replicata giornalmente la tabella. È usato nel caso di tabella con dimensioni contenute e una bassa variabilità.

3.2.2 Storicizzazione

Per evitare una perdita dei dati e continuare a garantirne l'integrità e l'accessibilità è necessario avere una storicizzazione del dato.
Esistono due opzioni differenti:

- **Creazione di tabelle *ombra* (HIS)**. Rappresentano una copia delle tabelle di DLT ma sono partizionate per JOB_ID (l'identificatore dell'estrazione) e vengono gestite con una operazione di **insert** (mantenendo in questo modo una copia storica dei dati per tutti i JOB_ID), contrariamente alle tabelle delta che non sono partizionate e sono gestite con operazioni di **truncate/insert** (ovvero vengono prima eliminati i dati precedenti e poi inseriti quelli nuovi corrispondenti al JOB_ID attuale)
- **Storico diretto sulle DLT** con partizionamento sempre JOB_ID

Al fine di garantire uniformità nell'architettura delle tabelle è possibile scegliere solo una delle due modalità descritte in base alle specifiche esigenze di ciascun progetto.

Nel progetto di tesi non saranno utilizzate né le tabelle di delta né quelle di storicizzazione. Questa scelta deriva dal fatto che i dati utilizzati per l'analisi derivano da file Excel/csv, che non essendo collegati ad una sorgente con costante aggiornamento giornaliero o mensile non offrono la possibilità di essere ricalcolati.

3.2.3 Modello multidimensionale

Per la progettazione dei sistemi informativi relazionali viene usato il modello E-R, modello entità-relazione. Esso permette di rappresentare graficamente la struttura delle informazioni e le relazioni tra gli elementi dei dati. Tuttavia, questo modello non è adatto per analizzare ed esprimere in modo accurato grandi moli di dati. (Adamson, 2010)
Nel caso di sistemi di supporto alle decisioni, che coinvolgono l'analisi di una vasta quantità di dati è necessario invece utilizzare il modello multidimensionale.
Con il modello multidimensionale (*Dimensional Fact Model*) è possibile organizzare i dati in una struttura a cubo. Si basa su quattro concetti chiave:

- Fatto: modella una specifica area di business (come ad esempio le vendite, gli ordini, la produzione). È caratterizzato da una o più misure.

- Misura: è l'aspetto quantitativo del fatto. Attraverso le misure si stabiliscono i KPI che guidano le imprese nelle proprie strategie di business. Ne sono esempi le quantità prodotte, il prezzo e così via.
- Dimensione: sono le coordinate di analisi del fatto. Tra questi abbiamo ad esempio data, prodotto, negozio.
- Attributo dimensionale: rappresentano le caratteristiche che descrivono una dimensione. Ad esempio, se la dimensione è "prodotto" gli attributi possono essere nome prodotto, categoria, fornitore, prezzo e così via.

Per organizzare i dati esistono differenti operazioni. Si ha: *pivoting*, ha lo scopo di modificare il punto di vista da cui analizzare i dati grazie alla rotazione degli assi del cubo; *slice & dice*, la prima permette di estrarre uno specifico sottoinsieme di dati all'interno del cubo, il secondo invece consente di fare una estrazione in base a criteri multipli; *roll-up & drill-down* per spostarsi all'interno di una gerarchia delle dimensioni, in particolare con roll-up si sale di livello gerarchico al fine di aggregare i dati ad un livello superiore (fornisce quindi una visione più aggregata), mentre con drill-down si scenderà verso livelli di dettaglio inferiori (fornendo quindi una visione più dettagliata). [27]
Prima di descrivere in modo più approfondito i livelli introdotti in precedenza, nel prossimo paragrafo verrà spiegata la metodologia e i criteri di selezione utilizzati per identificare i dati ritenuti utili per la mia analisi. Essendo i dati aziendali già disponibili al livello L2, il processo ETL verrà eseguito esclusivamente sui dati provenienti da fonti open.

3.3 Data Discovery

Quando si parla di Data Discovery ci si riferisce ad un processo fondamentale nel campo dell'analisi dei dati che coinvolge la raccolta, l'esplorazione e la valutazione dei dati provenienti da varie fonti al fine di comprendere e identificare eventuali tendenze, modelli o relazioni nascoste in essi.
Nel mio progetto ha avuto un ruolo di primaria importanza la ricerca di un set di informazioni completo e rappresentativo per essere di supporto alle successive valutazioni.
Avendo come principale obiettivo approfondire nel dettaglio le necessità e le condizioni del territorio italiano, in modo da delineare un panorama completo dell'Italia, sia in termini economici che in termini demografici, mi sono posta delle domande che potessero risultare utili nella scelta dei dati da raccogliere.
In particolar modo, in termini di analisi demografica mi sono interrogata su:

- Quali sono le caratteristiche demografiche, come ad esempio il reddito, l'età, l'occupazione, l'istruzione, della zona di interesse
- Quali sono i principali gruppi di consumatori
- Qual è il potenziale di crescita della zona

Mentre, in termini di concorrenza:

- Quali sono i principali competitor nella zona di interesse
- Quali sono i loro punti di forza e di debolezza
- Quali sono le loro posizioni geografiche

Per rispondere a queste domande ho individuato i seguenti indicatori rilevanti e utili per lo studio:

- Densità demografica
- Tasso di disoccupazione
- Incidenza della povertà relativa familiare
- Reddito medio mensile familiare
- Spesa media mensile familiare
- PIL pro-capite

Individuati questi fattori ho condotto la ricerca vera e propria dei dati. Per ottenere un'analisi completa ed affidabile è fondamentale l'integrità delle informazioni e la completezza di questi; pertanto, è stato necessario fare una selezione tra i vari siti e fonti a mia disposizione, quali ad esempio ISTAT, Aida, Statista. Dopo numerose verifiche ho ritenuto opportuno utilizzare i dati provenienti dall' ISTAT, ovvero l'ente nazionale responsabile della produzione e diffusione di dati statistici ufficiali sul territorio italiano, e quelli del MEF (Ministero dell'economia e delle Finanze).
I dati selezionati dall'Istituto Nazionale di Statistica sono quelli riguardanti:

- popolazione residente
- tasso di disoccupazione giovanile, che comprende la classe d'età dai 15 ai 34 anni
- tasso di disoccupazione totale
- spesa media mensile familiare per beni e servizi alimentari
- spesa media mensile familiare in totale
- pil pro-capite

ISTAT fornisce tali dati attraverso file nel formato csv, con valori disaggregati per regione. Solamente i primi tre indicatori offrono anche il dettaglio per provincia.
Tuttavia, i dati ISTAT non sono aggiornati all'anno in corso, ma presentano un orizzonte temporale che va dal 2004 al 2021. Essendo l'obiettivo di tale tesi fare una previsione sulle zone migliori in cui aprire nuovi punti di vendita risulta necessario ottenere una prospettiva futura. A tal proposito, inizialmente, ho pensato di fare un forecast utilizzando i dati degli ultimi 10 anni così da poter far

una previsione del 2022; successivamente, avendo trovato ricerche di settore più specifiche ho optato per un approccio differente utilizzando i dati ISTAT dell'anno 2021 e le pubblicazioni fornite dalla Banca d'Italia per realizzare un outlook più accurato e specifico. La Banca d'Italia, banca centrale italiana, infatti, pubblica regolarmente rapporti e analisi sull'economia italiana.
In particolare, dopo una attenta ricerca ho trovato che è stato previsto:

- Un aumento del 4,1% del PIL in Italia nel 2022
- Una diminuzione di 0,3 punti percentuali , così da raggiungere il 9,6 per cento, del tasso di disoccupazione
- Un aumento della quota di popolazione a rischio di povertà dell'1,3%
- Un deciso aumento dei consumi delle famiglie residenti pari al 3,7%
- Un calo del reddito disponibile reale poco più dell'1,5%

Successivamente, ho condotto una ricerca più approfondita sulle regioni di interesse nel mio studio al fine di verificare se le previsioni trovate per l'Italia nel complesso fossero rispecchiate anche a livello regionale. Ho quindi notato, sempre grazie all'ausilio di fonti quali ISTAT, Statista e Banca D'Italia, che le previsioni regionali sono coerenti ed allineate con le tendenze nazionali precedentemente identificate.
Per quanto riguarda le informazioni sul reddito medio delle famiglie sono stati utilizzati invece i dati forniti, sottoforma di Excel, dal Ministero dell'Economia e delle Finanze.
L'analisi relativa ai competitor si è rilevata più complessa poiché non è stato semplice individuare informazioni sulle posizioni geografiche dei competitor. L'unica fonte che ha fornito maggiori informazioni a riguardo è risultata essere il sito web delle Pagine Gialle. Questa pagina web offre informazioni su nome, numero di telefono, via, comune, provincia, regione, recensione e descrizione dei supermercati in questione. Per poter estrarre solo le informazioni rilevanti per la mia analisi (quindi nome supermercato, via, comune, provincia e regione) ho utilizzato un programma chiamato Octoparse. Octoparse è un software di web scraping che consente, grazie alla sua interfaccia intuitiva e alle potenti funzionalità, l'estrazione selettiva dei dati da pagine web strutturate. Al fine di garantire una maggiore precisione è stata infine svolta un'ulteriore operazione: sono state integrate le coordinate geografiche (longitudine e latitudine) dei punti vendita dei competitor.
Per svolgere tale operazione è stato utilizzato un algoritmo di programmazione in linguaggio Python.

3.4 Livello L0: DATA INGESTION

Questo livello è rappresentato dall'area di staging, fase iniziale in cui avviene tutto lo scarico delle informazioni dai sistemi sorgenti. Esistono varie tipologie di sistemi sorgente, i più frequenti sono ad esempio i sistemi operazionali su database (ovvero sistemi software, quali Oracle, SQL server, che gestiscono e archiviano i dati dell'organizzazione al fine di supportare tutte le operazioni quotidiane dell'azienda) oppure i file prodotti da fornitori.

Le modalità di lettura sono:

- Da tabella: vengono letti i dati giornalieri tramite rete per poi replicarli per intero all'interno del DWH
- Da file: vengono lette le informazioni contenute nell'estrazione giornaliera

Il processo di acquisizione e importazione dati provenienti da sorgenti differenti, usati per essere archiviati in un database o in altri sistemi di gestione dei dati, prende il nome di *data ingestion*. I dati in questione possono essere sia trasmessi in streaming in real time oppure ingeriti in lotti:

- Real time ingestion: utilizzato quando è necessario acquisire uno streaming di dati continuativo che non prevede un raggruppamento in batch per poter procedere alle altre fasi. Questo consente un processamento molto veloce.
- Batch processing: è più semplice da gestire rispetto al real time, poiché i dati vengono raccolti e processati in blocchi, per poi essere inviati periodicamente al sistema di destinazione per l'archiviazione. [28]

Tuttavia, nel caso in cui i dati sono in grandi quantità e in formati diversi risulta complicato per le aziende l'acquisizione a una velocità ragionevole. A questo scopo, vengono forniti programmi software personalizzati per ogni specifico ambiente di elaborazione. Questi permettono quindi l'automazione del processo di importazione dei dati e possono anche includere funzionalità di preparazione dei dati per strutturarli e organizzarli in modo efficiente per essere analizzati dalla BI e dalla BA (business analytics).

In questo livello tutte le tabelle create saranno precedute dal prefisso "STG" (staging area) e conterranno tutti i dati importati dello schema sorgente senza alcuna modifica.

3.4.1 Metadati

Per fornire informazioni complete sulle sorgenti, sul valore, sulle funzioni e sull'utilizzo dei dati memorizzati nel DWH si utilizzano i metadati. Con il termine metadato, vengono indicati tutti i dati usati per descrivere altri dati.

In base all'utilizzo è possibile distinguere due differenti categorie:

- Metadati di struttura (interni): sono di interesse per l'amministratore. Si riferiscono a tutte le informazioni sullo schema e sulla struttura dei dati, ne sono esempio le sorgenti, le trasformazioni, i vincoli ecc.
- Metadati di contenuto (esterni): sono di interesse per gli utenti. Si riferiscono a tutte le informazioni aggiuntive sui dati che arrivano da fonti esterne, come il significato, le definizioni, le unità di misura, le regole di calcolo e così via.

Quindi per riassumere i primi definiscono l'architettura i secondi descrivono l'informazione. (Chiarello, 2020) [29]

È anche possibile fare una classificazione in base al contesto in cui vengono considerati, distinguendoli in metadati globali e di processo. Quelli globali forniscono informazioni a livello di sistema e coprono tutti i processi, offrendo così una visione più ampia delle attività; i secondi sono i metadati specifici per ogni singolo processo o sistema, forniscono quindi informazioni dettagliate sulle varie operazioni coinvolte.

Solitamente, per garantire l'accessibilità da parte dei vari sistemi i metadati vengono registrati all'interno di una tabella. Fornendo in modo preciso le informazioni sullo stato di un processo, evitano la duplicazione delle istanze o l'avvio in momenti scorretti, consentendo così una gestione ottimale del DWH, garantendo una corretta estrazione e trasformazione dei dati.

Dopo questa prima parte maggiormente teorica su cosa sono i metadati e su come questi giochino un ruolo di cruciale importanza nella descrizione dei dati e nel processo di ETL, passiamo adesso alla trattazione dell'implementazione e della creazione del Data Mart tramite l'utilizzo di Data Stage. Bisogna però specificare che, in fase di prototipo, il cliente non ha fornito una macchina performante; quindi, si è adottata la tecnica *push down.* Con questa tecnica alcune trasformazioni verranno eseguite all'interno del database, diventando così tecnicamente un *T-E-L.*

Tramite le varie funzionalità del software utilizzato è stato possibile importare diversi tipi di file così da costruire il nuovo Data Mart in SQL Server *Supermercato.PROSPECT*. In questa prima fase verranno create le tabelle STG contenenti i dati importati da ISTAT e MEF.

Nella tabella seguente vengono elencate le tabelle create in questo livello specificando la tipologia dei file sorgente.

DATABASE TABLE	METADATA TYPE	METADATA NAME
STG_istat_densita_popolazione	File: .csv	Istat densità popolazione
STG_istat_pil_procapite_prezzi_correnti	File: .csv	Istat pil pro-capite prezzi correnti
STG_istat_poverta_regione	File: .csv	Istat povertà regione
STG_istat_tasso_disoccupazione_giov	File: .csv	Istat tasso disoccupazione giovanile
STG_ istat_tasso_disoccupazione_tot	File: .csv	Istat tasso disoccupazione totale
STG_istat_voce_spesa_alimenatari	File: .csv	Istat voce spesa alimentari
STG_istat_voce_spesa_totale	File: .csv	Istat voce spesa totale
STG_mef_reddito_medio_famiglie	File: .Excel	Reddito medio famiglie

DataStage offre diverse opzioni per importare i file. Nel caso di file CSV (Coma Separate Values), viene utilizzato uno stage chiamato *Sequential File,* che consente la lettura dei dati da file sequenziali, inclusi file CSV. Il Sequential File rappresenterà i file sorgente (ovvero i vari file csv provenienti dai siti ISTAT e MEF). Utilizzeremo invece un ODBC connector come tabella di destinazione, che corrisponderà alla tabella STG.

Prima di importare un file (figura 12) è necessario creare un progetto e successivamente creare un job, che rappresenta il flusso logico delle attività di elaborazione dei dati.

Una volta creato il job, bisogna aggiungere il Sequential File, che ci permetterà di specificare tutti i dettagli del file csv che si desidera importare. Per fare ciò, è necessario trascinare lo stage dalla palette e poi successivamente configurarlo, specificando la sorgente del file, il formato dei dati, il delimitatore di campo (nel nostro caso sarà la virgola), la tipologia della prima riga del file (se contiene dati o nomi delle colonne) e così via.

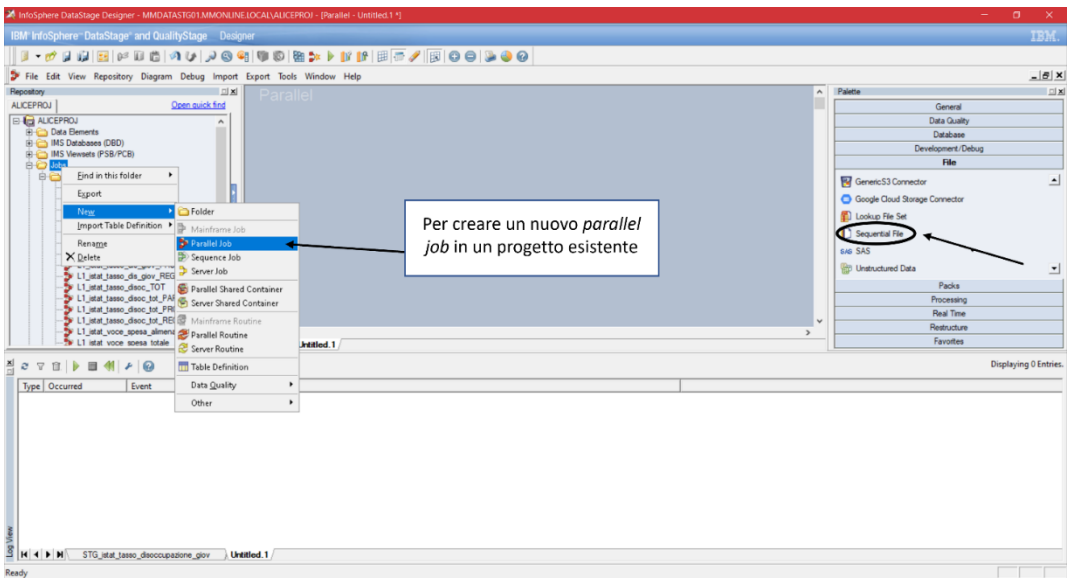

Figura 11: Creazione nuovo job

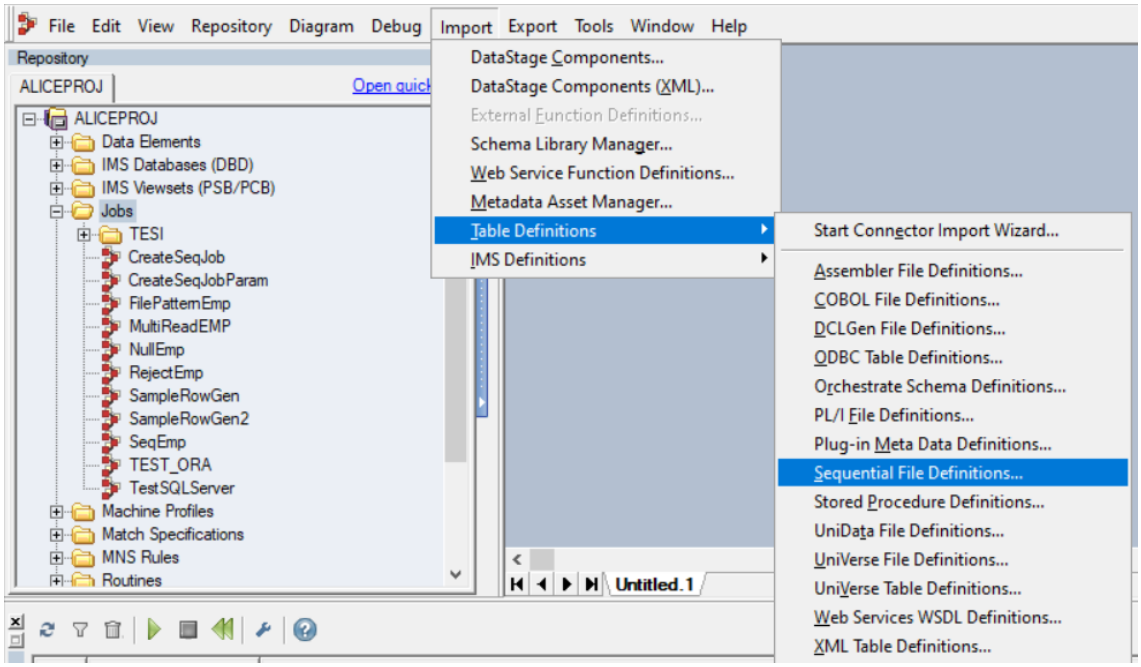

Figura 12: Importazione dei file

Passiamo ora alla tabella di destinazione. Questa tabella è collegata con il database in cui sono già presenti i dati dell'azienda in esame e in cui è stata precedentemente creata la struttura delle varie tabelle STG.

Prima di tutto, è necessario connettersi al Database, in modo da poter importare (come fatto per il Sequential File, ma questa volta bisogna selezionare ODBC Table definitions) le tabelle create tramite SQL Server su DBeaver (la piattaforma di amministrazione del database).

Aggiungiamo quindi al job, tramite la palette, uno stage chiamato "*ODBC connector*", specificando tutti i dettagli della connessione.

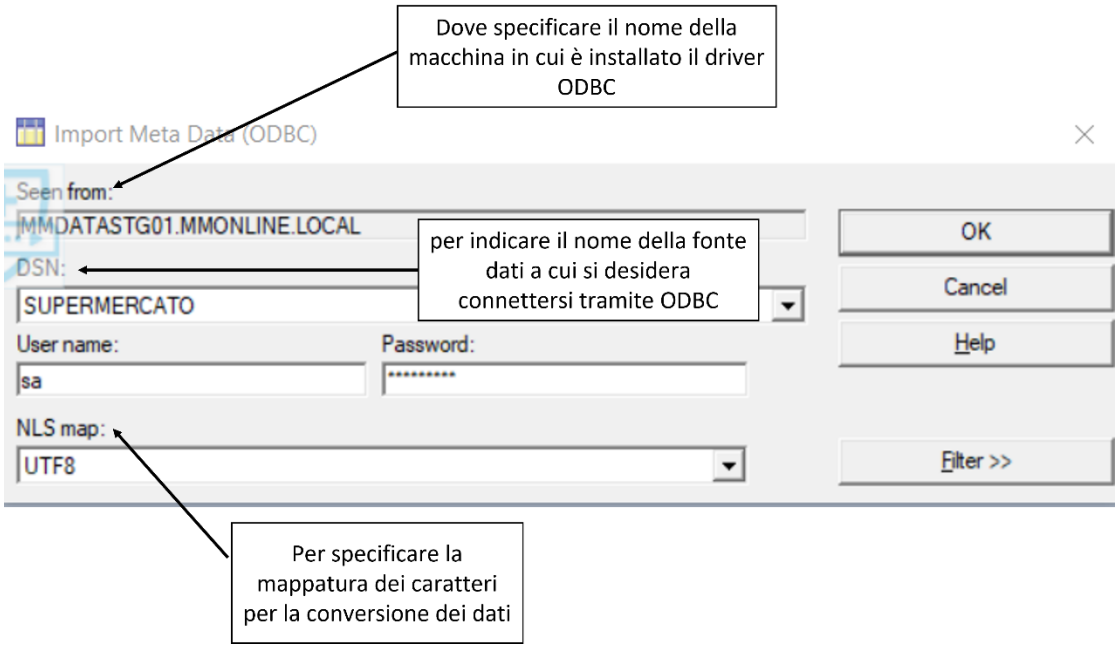

Figura 13: Connessione al database

Una volta stabilita la connessione, bisogna specificare le altre proprietà, come la tipologia di *write mode* (nel nostro caso, "*insert then update*" che consiste nell'inserimento dei nuovi dati e nell'aggiornamento di quelli esistenti), se generare o meno SQL (in questo caso viene impostato su *yes* e in *table name* verrà inserito il file di origine dei dati, come ad esempio nel caso della tabella STG_istat_tasso_disoccupazione_giovanile metteremo PROSPECT.STG_istat_tasso_disoccupazione_giovanile dove Prospect è lo schema in cui si trovano le tabelle). Infine, in *column* verrà effettuato il *load* (caricamento) della tabella.

Prima di popolare le tabelle di staging, in questa fase si è deciso di utilizzare un altro stage chiamato *Transformer*, che si trova nella sezione di processing della palette. L'utilizzo di questo stage permette di uniformare il formato dei dati, consentendo di cambiare sia il datatype dalla sorgente verso la destinazione, sia di fare trasformazioni al dato. In particolare, sono stati eliminati gli spazi di ciascun valore (utilizzando la funzione '*trim*') e sono stati convertiti tutti i dati in maiuscolo (grazie alla funzione *"UpCase"*).

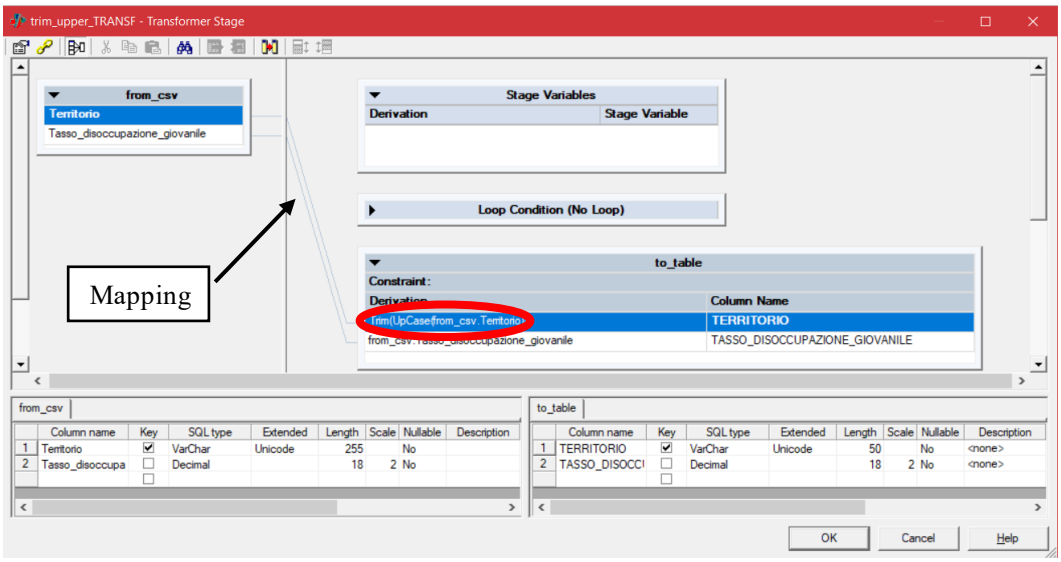

Figura 14: Trasformer stage

Nella parte superiore viene visualizzato il cosiddetto *mapping,* ovvero l'associazione tra i campi di input (in questo caso from_csv, ovvero il link collegato al Sequential File) e i campi di output (to_table, corrispondente all'ODBC connector). Serve quindi a definire come fluiscono i dati da una sorgente ai campi di destinazione all'interno di un processo.

Qui di seguito vediamo la rappresentazione del job usato per il popolamento della tabella STG_istat_tasso_disoccupazione_giovanile, ma di uguale struttura saranno anche tutti gli altri job per le altre STG. É importante sottolineare che, al fine di prevenire errori durante il caricamento dei dati, ogni file occuperà un Job diverso.

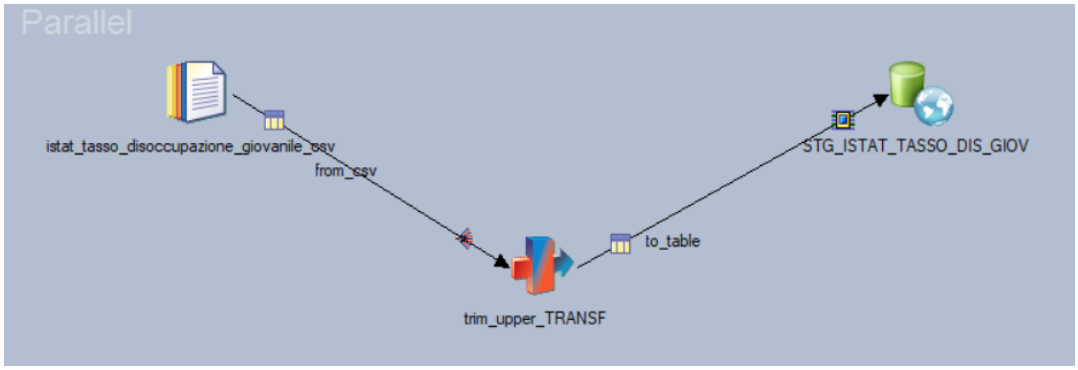

Figura 15: Popolamento STG_istat_tasso_disoccupazione_giovanile

3.5 Livello L1: DATA OPERATION

Il livello L1 è indubbiamente il fulcro centrale dell'intero processo di ETL. Esso comprende i processi di data Quality, normalizzazione dei dati e tutte le trasformazioni dei file sorgente.

In questa fase i dati verranno estratti solo dalle tabelle create nell'area di staging. In seguito all'estrazione, i dati vengono elaborati, trasformati e caricati nel Database *Supermercato.PROSPECT*, ma in modo normalizzato al fine di migliorare il processo e di avere un controllo continuo su tutte le attività.

3.5.1 Data Quality

La Data Quality "*consiste nella pianificazione, implementazione e controllo delle attività che applicano tecniche di gestione della qualità dei dati, al fine di garantire che siano adatti allo scopo e soddisfino le esigenze degli utilizzatori*". (DAMA, Global Data Management Community)

Abbiamo detto più volte quanto i dati risultino essere di fondamentale importanza nel processo decisionale di una azienda, influenzando non solo le scelte odierne dell'organizzazione ma anche tutte quelle future. A tal proposito, negli anni l'attenzione sulla qualità dei dati ha ottenuto sempre maggior rilievo e considerazione, poiché, come affermato dalla società di consulenza Gartner, "*la scarsa qualità dei dati distrugge il valore del business*" (Saul, Alan, Melody, & Ted, 2020)

La qualità dei dati è determinata da vari fattori, quali:

- Completezza, quanti dati sono stati archiviati rispetto a quanto era previsto
- Unicità, nessuna istanza deve essere registrata più di una volta nel sistema in modo da evitare ridondanze
- Tempestività, indica il grado in cui i dati riflettono la realtà nel momento in cui sono necessari per il processo decisionale
- Validità, riguarda la conformità dei dati alla sintassi o al formato definito

- Accuratezza, quanto i dati descrivono con precisione l'oggetto o l'evento che intendono rappresentare
- Consistenza, è l'assenza di differenze o discrepanze quando vengono confrontate due o più rappresentazioni di un determinato dato o concetto (DAMA UK, 2013) [30]

Durante la progettazione del Data Mart diverse sono state le operazioni di Data Quality eseguite. Un esempio di tali modifiche è la rimozione degli spazi prima o dopo il valore; oppure la rimozione dei caratteri speciali come il trattino ("- "), ad esempio nella stringa "Trentino-Alto-Adige" che diventa "Trentino Alto Adige". Sono state anche attuate modifiche nella dimensione dei campi così da evitare la *data truncation.*

Vediamo adesso nel dettaglio come sono state create e popolate le tabelle L1 tramite DataStage.

Prima di tutto, è importante precisare che non tutti i dati ISTAT e MEF includevano informazioni dettagliate sia sulle province che sulle regioni. Per risolvere questo problema è stato quindi formulata una ipotesi: nel caso in cui mancasse il dettaglio della provincia si è ipotizzato che il dato equivalesse a quello della regione.

Durante questo livello è stato fondamentale definire le chiavi identificative, così da poter stabilire un collegamento tra i dati open e quelli dell'azienda stessa. Infatti, i dati provenienti dalle fonti open contenevano solo colonne come "territorio" (che potevano includere regioni e/o provincie) e il valore ad esso associato.

Come primo passo sono state create le strutture delle varie tabelle L1 tramite SQL server su DBeaver, inserendo come colonne regione_code, regione_desc, provincia_code, provincia_desc, il valore ad esse associato, Ins time e Upd time. È stato in seguito utilizzato DataStage per il popolamento di tali colonne. In particolare, questo è avvenuto grazie all'unione delle tabelle STG con una tabella proveniente dal database aziendale, la L2.DIM_GEO, che conteneva al suo interno già le corrispondenze regione/provincia con il codice identificativo.

Nel caso in cui nella colonna *territorio* delle tabelle STG erano presenti sia i nomi delle regioni che delle province è stato necessario creare tre diverse tabelle (nel caso del tasso di disoccupazione giovanile, ad esempio, si ha L1_istat_tasso_dis_giov_REG, L1_istat_tasso_dis_giov_PRO, L1_istat_tasso_dis_giov) e tre job differenti, uno per popolare la tabella con le informazioni sulla regione, uno per le informazioni sulle provincie e uno per l'unione dei due.

Iniziamo con il popolamento della tabella L1_istat_tasso_dis_giov_PROV.

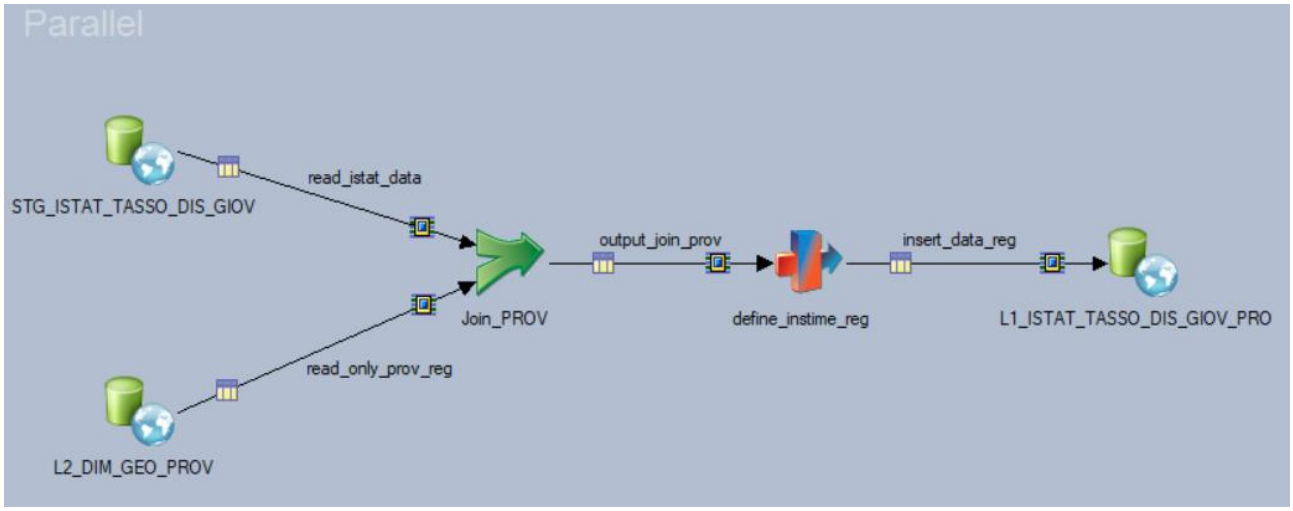

Figura 16:Popolamento L1_istat_tasso_dis_giov_PRO

Come possiamo vedere dalla figura 16, si hanno due tabelle sorgenti rappresentati dai due ODBC connector:

- Il primo rappresenta la tabella STG, che contiene tutti i dati relativi al tasso di disoccupazione giovanile. La configurazione di tale stage viene fatta come già visto in precedenza.
- Il secondo è L2_DIM_GEO. Questa è una tabella proveniente dal database dell'azienda. In questo caso non avremmo bisogno di tutte le colonne presenti nella tabella ma solo di quelle necessarie per la nostra L1. Per questo motivo nelle proprietà di questo specifico ODBC connector viene selezionato *no* in "*generate SQL*" e verrà scritta la seguente query:

select distinct TRIM(UPPER(dg.PROVINCIA_DESC)) **as** TERRITORIO,
TRIM(UPPER(dg.PROVINCIA_CODE)) **as** PROVINCIA_CODE,
TRIM(UPPER(dg.REGIONE_DESC)) **as** REGIONE_DESC,
TRIM(UPPER(dg.REGIONE_CODE)) **as** REGIONE_CODE
from prospect.DIM_GEO dg;

che ha lo scopo di selezionare dalla tabella DIM_GEO solo le colonne provincia_desc, provincia_code, regione_desc e regione_code. Il termine "*as*" viene utilizzato per assegnare degli alias alle colonne. *TRIM* è una funzione che serve ad eliminare gli spazi, mentre *UPPER* per scrivere tutto in maiuscolo.

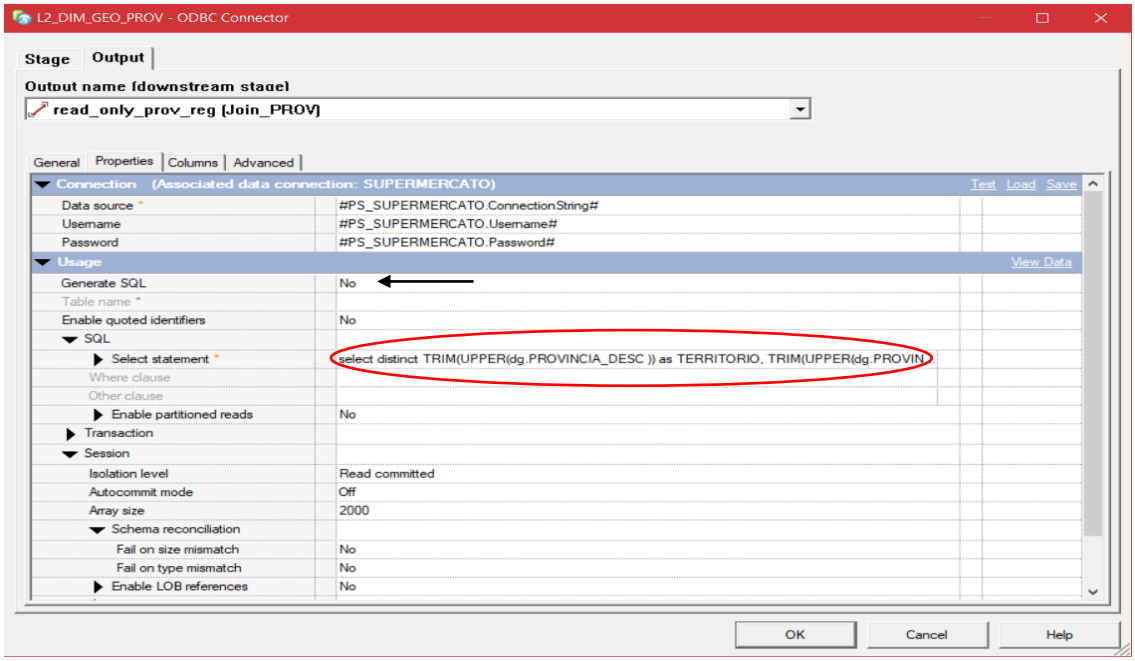

Figura 17: L2_DIM_GEO_PROV

I link uscenti dai due ODBC convergono in un nuovo stage: il JOIN. Questo stage fa parte del gruppo *processing* della palette. È utilizzato per combinare o unire due o più set di dati in base a una o più chiavi di join specificate. Esistono varie tipologie di join:

- Inner join: verranno restituiti esclusivamente i record che trovano una corrispondenza tra le tabelle.
- Left outer join: vengono restituiti tutti i record della tabella di sinistra (ovvero la prima tabella specificata nella join) e solo i record corrispondenti dalla tabella di destra (la seconda specificata).
- Right outer join: vengono restituiti tutti i record della tabella di destra (ovvero la seconda tabella specificata nella join) e solo i record corrispondenti dalla tabella di sinistra (la prima tabella).
- Full outer join: vengono restituiti tutti i record, è la combinazione di right e left outer join.

In DataStage, nella schermata del join si distinguono tre componenti principali: stage, input e output.

Nello stage è possibile configurare i parametri del join, come la specifica della chiave e il tipo di join. Per questa analisi è stato utilizzato l'inner join, che ha così consentito di individuare e correlare i corrispettivi tra le due tabelle coinvolte, scartando invece tutti i valori che non hanno trovato alcuna corrispondenza.

L'attributo scelto come chiave è TERRITORIO (per la L2_DIM_GEO, essendo nel caso del L1 per provincia, questo corrisponde alla provincia_desc).

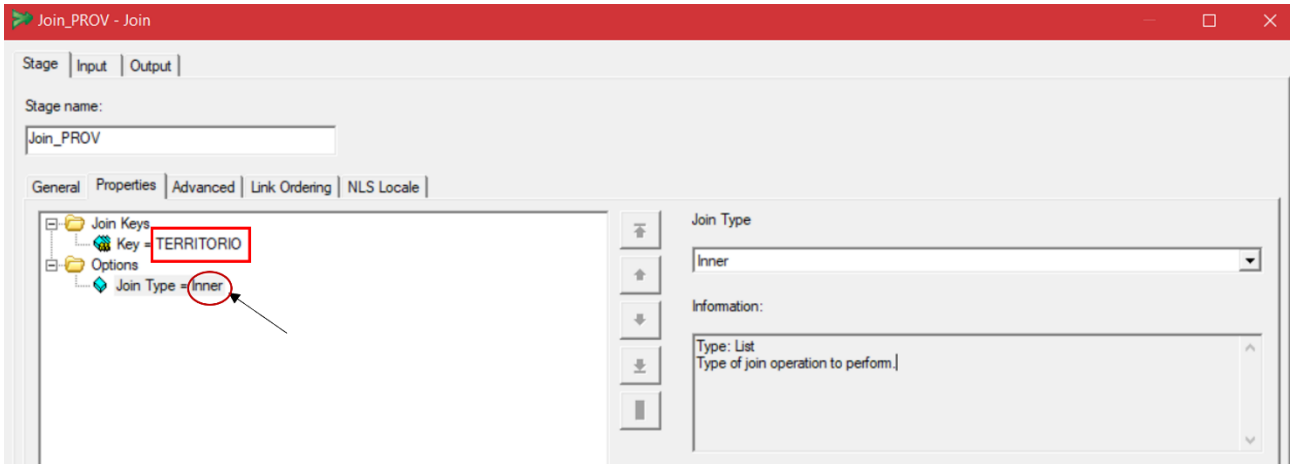

Figura 18: Stage Join

Nell'input vengono rappresentate entrambe le tabelle sorgente con le corrispettive colonne. In questa parte è stato possibile modificare il datatype e la *lenght* delle colonna Territorio in modo da essere uniformi tra le due tabelle.

Lato output troviamo invece il risultato del join. Per popolare la tabella di destinazione è necessario fare il mapping, associando tutte le colonne che hanno il medesimo significato.

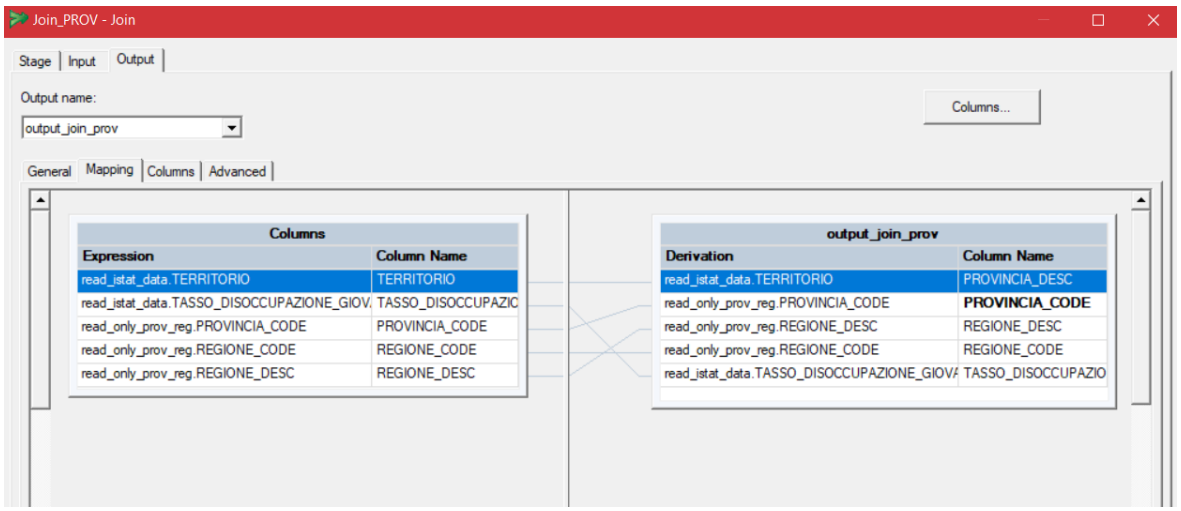

Figura 19: Mapping output

È stato inoltre inserito tra il join e la tabella di destinazione uno stage Transformer in modo da poter generare il valore di INS time e UPD time. Questi valori servono a tenere traccia della cronologia delle modifiche apportate ai dati nel database. In particolare, *Ins time* (insertion time) indica il momento in cui viene inserito per la prima volta un record nel database, mentre *Upd time* (update

time) indica l'ultima volta in cui il dato è stato aggiornato. La funzione usata nel Transformer è il *CurrentTimestamp*.
Anche in questo caso la tabella di destinazione è rappresentata da un ODBC connector che dovrà essere configurata come quelle già viste in precedenza.

Lo stesso procedimento verrà fatto anche per il secondo job per popolare L1_istat_tasso_dis_giov_REG. Le uniche differenze sono:

- La query del L2_DIM_GEO è :
  select distinct TRIM(UPPER(dg.REGIONE_DESC)) **as** TERRITORIO,
  TRIM(UPPER(dg.REGIONE_CODE)) **as** REGIONE_CODE
  from prospect.DIM_GEO dg;
  Notiamo infatti che non verranno considerate le colonne relative alla provincia.
- La chiave del join è sempre territorio, ma per la L2_DIM_GEO non corrisponde più alla provincia_desc ma alla regione_desc.

Con questi due job sono state quindi popolate le seguenti tabelle:

- L1_istat_tasso_dis_giov_REG che presenta le seguenti colonne: REGIONE_DESC, REGIONE_CODE, TASSO_DISOCCUPAZIONE_GIOV_REG, INS_TIME, UDP_TIME
- L1_istat_tasso_dis_giov_PRO con seguenti colonne: PROVINCIA_DESC, PROVINCIA_CODE, REGIONE_DESC, REGIONE_CODE, TASSO_DISOCCUPAZIONE_GIOV_PRO, INS_TIME, UDP_TIME.

Vediamo adesso l'ultimo job, necessario per il popolamento della tabella finale**:** **L1_istat_tasso_dis_giov** contenete le seguenti colonne: PROVINCIA_DESC, PROVINCIA_CODE, REGIONE_DESC, REGIONE_CODE, TASSO_DISOCCUPAZIONE_GIOV_REG, TASSO_DISOCCUOAZIONE_GIOV_PROV, INS_TIME, UDP_TIME.

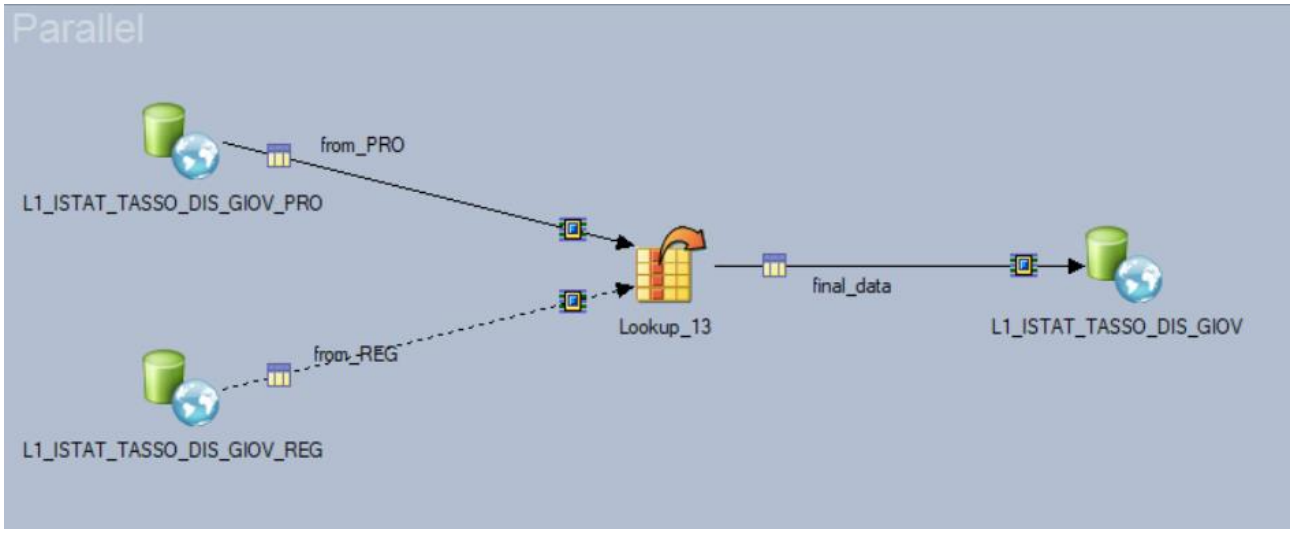

Figura 20: Job per popolare L1_istat_tasso_dis_giov

Le due tabelle sorgenti in questo caso sono le L1 per provincia e per regione create nei due job precedenti, la tabella di destinazione è invece L1 finale.

Nel flusso di lavoro sono quindi presenti tre ODBC, due come sorgenti e uno di destinazione. Tra sorgenti e destinazione è stato inserito un nuovo stage: il *Lookup.* Questo stage è utilizzato per combinare i flussi di dati e recuperare eventualmente attributi aggiuntivi associati ad un determinato record. Serve quindi per arricchire i record in base a determinati criteri specifici.

La tabella di riferimento, che prende il nome di tabella master, nel nostro caso è quella della provincia. È la tabella contenente dati aggiuntivi rispetto alla tabella di input e presenta una chiave primaria (chiave di lookup) che servirà nella ricerca dei valori corrispondenti nella tabella di input. In questo caso specifico la chiave è rappresentata dal campo "regione_code".

Quando lo stage verrà eseguito, il sistema confronterà i valori della chiave con i record presenti nella tabella di input con quelli della tabella master. Nel caso in cui viene trovata una corrispondenza, verranno recuperati i dati aggiuntivi e successivamente verranno aggiunti al record corrispondente nella tabella di input.

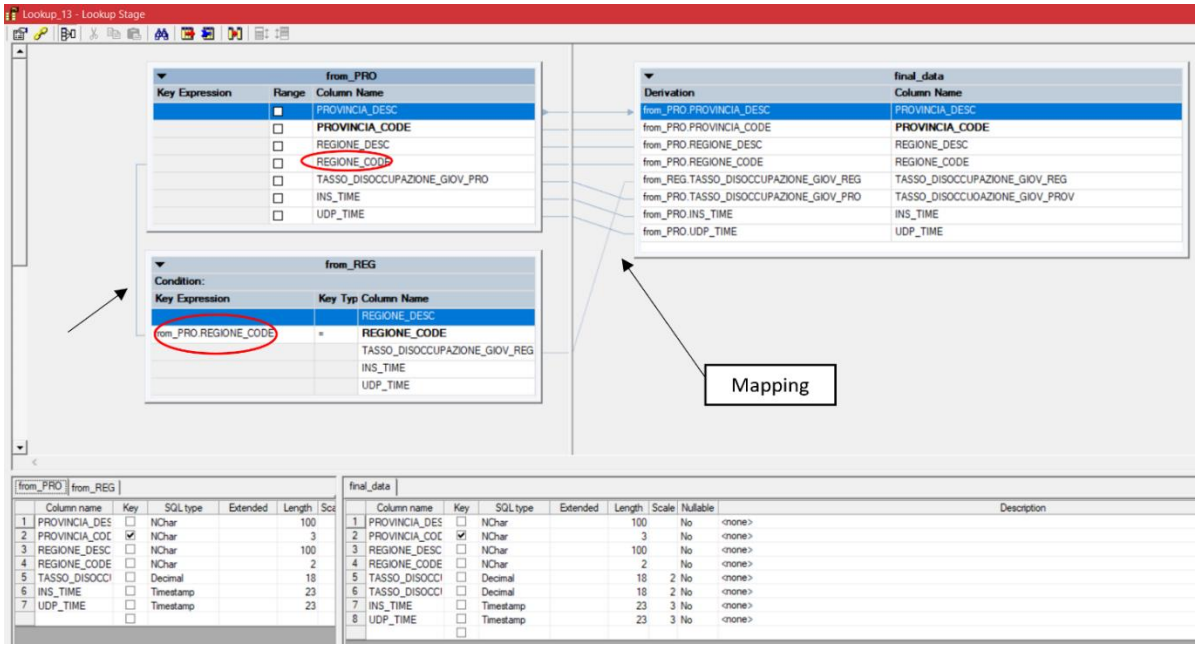

Figura 21: Lookup

Per la creazione delle altre tabelle L1 vengono svolti i medesimi procedimenti visti per la tabella L1_istat_tasso_dis_giov. Nel caso in cui nella colonna *Territorio* delle STG sono presenti solo i nomi delle regioni non è necessario creare tre differenti job, ma tutto il processo verrà svolto all'interno di un solo job:

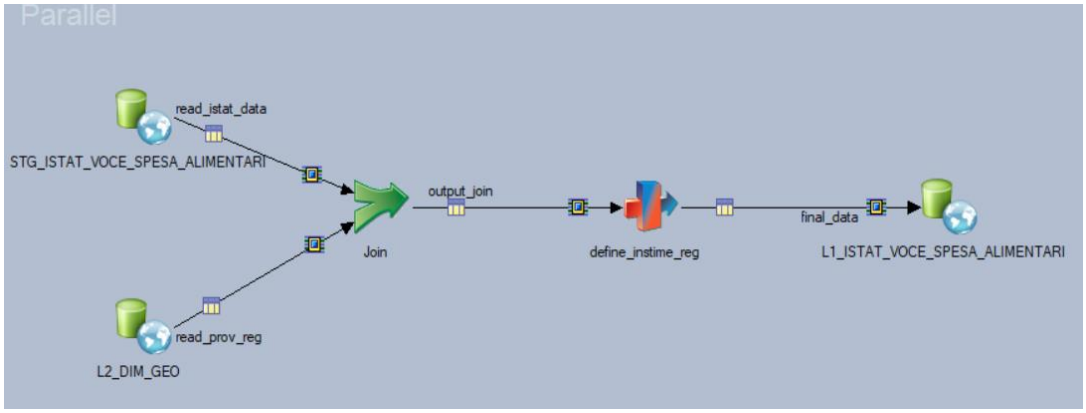

Figura 22: Popolamento L1_istat_voce_spesa_alimentari

In questi casi la query nel L2_DIM_GEO sarà:

select distinct TRIM(UPPER(dg.PROVINCIA_DESC)) **as** PROVINCIA_DESC,
TRIM(UPPER(dg.PROVINCIA_CODE)) **as** PROVINCIA_CODE,
TRIM(UPPER(dg.REGIONE_DESC)) **as** TERRITORIO,
TRIM(UPPER(dg.REGIONE_CODE)) **as** REGIONE_CODE
from prospect.DIM_GEO dg

dalla quale notiamo che la chiave di join sarà sempre Territorio ma, nella L2_dim_geo corrisponderà al campo "regione_desc".

3.6 Livello L2

Per la rappresentazione logica e l'organizzazione dei dati esistono diversi modelli comunemente utilizzati con i database relazionali (sistemi di gestione dei dati che organizzano le informazioni in righe e colonne, che nel loro insieme formano una tabella, e attraverso l'utilizzo di chiavi primarie e chiavi esterne stabiliscono connessioni tra di esse). Tra i diversi modelli, quelli più utilizzati sono i già citati star schema e snowflake schema. In questo progetto, data la relativa semplicità delle relazioni tra le diverse tabelle, si è deciso di implementare uno star schema.

Lo star schema è composto da una tabella centrale, che prende il nome di *fact table* (tabella dei fatti), e da una serie di tabelle correlate con essa, le *dimension table* (tabella delle dimensioni). Il tipo di relazione tra la tabella dei fatti e quella delle dimensioni è uno-a-molti. Infatti, una tabella dei fatti può essere collegata a infinite tabelle delle dimensioni, mentre queste ultime sono legate solo con una fact e senza la possibilità di legarsi con altre dimension table.

Generalmente, nella fact table vengono memorizzate le misure numeriche del business (quali vendite, quantità, profitti) . Al suo interno esse contengono le colonne chiave di dimensione, ovvero tutte le chiavi esterne che la correlano con ciascuna delle tabelle delle dimensioni, permettendo di collegare ciascuna riga della fact con le righe corrispondenti delle dimension.

Le dimension table contengono invece le descrizioni funzionali della specifica dimensione del business, come ad esempio quella relativa al prodotto, al tempo e così via.

Le principali caratteristiche di questa tipologia di schema sono quindi:

- La presenza di una struttura semplice che permette così una comprensione facile;
- Query molto performanti, che riducono i join da effettuare nelle tabelle;
- Tempi di caricamento più lunghi a causa della de-normalizzazione della dimension table, che provoca una ridondanza dei dati e quindi un maggiore spazio di archiviazione della tabella;
- La fact table, contrariamente alla precedente, è in terza forma normale garantendo così l'integrità dei dati ed evitando la ridondanza; (Adamson, 2010)
- Questo schema è, inoltre, anche supportato da molteplici strumenti di business intelligence. [31]

Nel caso in esame, la tabella dei fatti è indicata con il nome FACT_OPEN, che sarà collegata a due tabelle delle dimensioni: la DIM_REGIONE e la DIM_PROVINCIA.

Vediamo strutturalmente quali sono i campi di ogni tabella.

FACT_OPEN			
COLUMN_NAME	**DATA_TYPE**	**PRIMARY_KEY**	**NULLABLE**
OPEN_SK	int	yes	no
INDICATORE	varchar	no	no
REGIONE_SK	int	no	no
PROVINCIA_SK	int	no	no
ANNO	int	no	no
PREDIZIONE_ANNO_FUTURO	float	no	no
MISURA	float	no	no

La *primary key (pk)* , rappresentata da "OPEN_SK" nella tabella dei fatti, è un attributo che ha il compito di identificare univocamente ogni riga presente nel database. L'acronimo "SK" indica la presenza di una *surrogate key* , ossia una chiave che viene creata artificialmente così da essere utilizzata come identificatore indipendente da tutte le altre informazioni presenti nella riga. L'utilizzo e la definizione di una chiave primaria comportano diversi benefici:

- Garanzia di unicità dei dati, la *pk* assicura che ogni riga della tabella abbia un valore univoco ad essa associata , garantendo così l'identificazione inequivocabile di ogni record;
- Impedisce la duplicazione dei record in una tabella, preservando così l'integrità dei dati e prevenendo eventuali inconsistenze;
- Aiuta a stabilire le relazioni con le altre tabelle; essa, infatti, può essere utilizzata come chiave esterna che collega le diverse tabelle tra di loro. Ad esempio, i campi "provincia_sk" e "regione_sk" rappresentano rispettivamente le chiave primarie della DIM_PROVINICIA e della DIM_REGIONE, mentre all'interno FACT_OPEN svolgono la funzione di chiavi esterne.
- Garantisce l'accessibilità a livello di riga, semplificando così le operazioni di ricerca, di aggiornamento ed eliminazione di dati specifici. [32]

La colonna "Indicatore" comprende i nome di tutti gli indicatori precedentemente definiti (tasso di disoccupazione giovanile e totale, incidenza della povertà, pil pro-capite, voce spesa alimentari e totale, reddito medio delle famiglie, densità della popolazione. I relativi valori, invece, sono inseriti nella colonna "Misura" per l'anno 2021, mentre ,nel caso dei valori ottenuti tramite previsione, vengono collocati nella colonna "Predizione anno futuro".
Le tabelle delle dimensioni presentano invece le seguenti strutture:

DIM_REGIONE			
COLUMN NAME	**DATA_TYPE**	**PRIMARY_KEY**	**NULLABLE**
REGIONE_SK	int	yes	no
REGIONE_CODE	varchar	no	no
REGIONE_DESC	varchar	no	no

DIM_PROVINCIA			
COLUMN NAME	**DATA_TYPE**	**PRIMARY_KEY**	**NULLABLE**
PROVINCIA_SK	int	yes	no
PROVINCA_CODE	varchar	no	no
PROVINCIA_DESC	varchar	no	no

Nella figura seguente vengono visualizzati i collegamenti tra le tabelle:

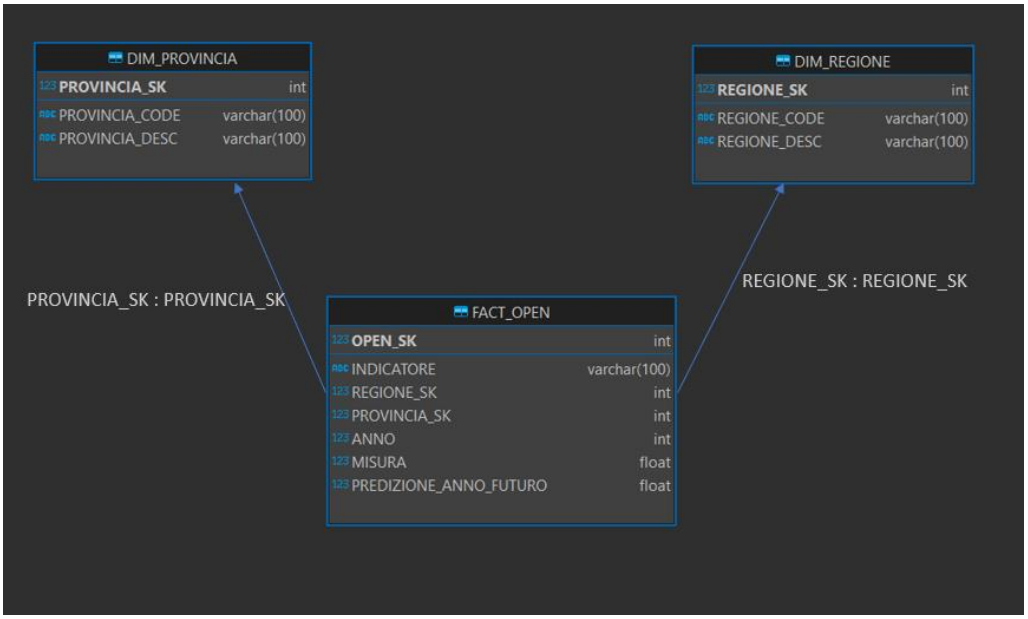

Figura 23: Star Schema FACT_OPEN

3.7 L2 Cross Analysis

Dopo aver completato il processo ETL dei dati provenienti da fonti aperte, si è finalmente in possesso di un data Mart integrato che include sia i dati dell'azienda che quelli provenienti da fonti aperte. Prima di spiegare nel dettaglio cosa si intende per cross analysis, è opportuno specificare i dati aziendali utilizzati.

Come specificato all'inizio di questo capitolo, non è stato necessario eseguire un processo ETL per i dati provenienti dal database aziendale, in quanto essi erano già al livello di L2.

All'interno del database aziendale, sono state selezionate specifiche tabelle contenti informazioni ritenute utili ai fini delle mie analisi:

- La tabella delle dimensioni DIM_GEO, che contiene tutte le informazioni geografiche, come codici e nomi delle regioni, delle province e dei comuni, le coordinate geografiche e così via;

- La tabella DIM_PDV, che fornisce informazioni dettagliate sui punti di vendita, tra cui nome del brand, codice azienda, indirizzo, località, data di apertura ed eventuale data di chiusura, stato del punto di vendita (attivo o meno);
- La tabella DIM_CALENDAR, che include tutte le informazioni temporali rilevanti;
- La tabella dei fatti FACT_VENDITE_TRASFERIMENTI, collegata con la dim_pdv tramite la chiave esterna "pdv_sk" e con la dim_calendar tramite "date_sk"

Prima di intraprendere l'analisi integrata, sono state inoltre condotte diverse analisi esclusive sui dati aziendali. Ad esempio, attraverso l'esecuzione di apposite query sono state identificate le regioni con maggior numero di punti vendita, evidenziando Basilicata, Campania e Abruzzo e in particolare le province di Potenza, Matera, Salerno, Avellino e L'Aquila. Di conseguenza, è stata presa la decisione di concentrare tutte le future analisi su queste specifiche province.
Successivamente, sono stati esaminati numeri di apertura e chiusura dei punti di vendita per ciascun anno, notando che gli anni 2016 e 2019, grazie ad un equilibrio tra chiusure e aperure sono risultati i migliori, mentre il 2017 e 2018 i peggiori. In termini di fatturato, calcolato anche esso tramite query specifiche, è stato invece osservato un costante incremento nel corso degli anni.

Dopo aver concluso tutte queste analisi preliminari è stato quindi possibile creare il data Mart. Esso contiene dati **normalizzati** provenienti da fonti diverse, sia aziendali che aperte, consentendo così di poter confrontare e analizzare i dati in modo uniforme al fine di ottenere informazioni significative per il processo decisionale dell'azienda.

La *cross analysis* è, infatti, un metodo che coinvolge simultaneamente fonti differenti al fine di ottenere una panoramica più dettagliata e ampia del contesto in cui opera l'azienda.
L'introduzione di tale attività all'interno dell'azienda mi ha permesso di fornire loro una metodologia che gli permetta di raggiungere un vero e proprio vantaggio competitivo rispetto ai competitor. Essi, spesso si affidano esclusivamente a strumenti ERP (Enterprise Resource Planning), strumenti per l'integrazione e la gestione dei dati interni dell'azienda.
Contrariamente agli ERP la cross analysis, infatti, permette di attingere ad una vasta gamma di dati, non solo interni ma anche esterni, rilevando in questo modo correlazioni e relazioni tra le diverse tipologie di dati e ottenendo informazioni, scoperte significative ed innovative per le aziende, risultando così anche più inclini al cambiamento. Infatti, grazie alle continue analisi, esse riescono a monitore e rispondere prontamente alle nuove tendenze che possono influenzare il contesto aziendale.

Nello specifico, per questo elaborato la cross analysis è risultata essenziale per l'identificazione delle zone più appropriate e profittevoli per aprire nuovi punti di vendita.

3. 8 Auditing ETL

Durante tutto il processo ETL è fondamentale monitorare e registrare le varie operazioni. Il processo di *auditing* ha, infatti, lo scopo di soddisfare i seguenti obiettivi:

- Controllare se sono presenti errori gravi
- Rilevare la presenza di eventuali anomalie nei dati
- Registrare e conservare una traccia delle modifiche apportate ai dati durante il processo di trasformazione

Per poter confermare che i dati caricati corrispondono a quelli previsti, tali processi dovrebbero includere:

- Un conteggio generale delle righe
- Totali aggregati (come gli import finanziari o altri dati di riepilogo)

In alcuni casi, potrebbe essere necessario attuare un controllo più approfondito oppure bisognerebbe esclusivamente verificare che i dati supportano determinati valori.

Infine, un'ulteriore verifica è il controllo dei casi in cui non è stato caricato alcun dato. Questo è un avvenimento che accade innumerevoli volte; infatti, anche se il processo ETL viene completato correttamente potrebbe succedere che le righe caricate risultino pari a zero. Diversi sono i motivi di tale risultato, come ad esempio il file di origine è vuoto oppure una query viene configurata in modo errato non restituendo così alcuna riga.

In definitiva, possiamo quindi dire che il processo di auditing aiuta a garantire integrità, accuratezza e tracciabilità dei dati.

3.9 Best geo-locations

Per svolgere una analisi completa ed ottenere un risultato affidabile è stato necessario creare diverse query, che interrogavano sia i dati aziendali che quelli provenienti dalle fonti aperte, con l'obiettivo finale di individuare quali tra le varie province italiane risultasse la più appetibile economicamente per l'apertura dei nuovi punti vendita della GDO.

Come precedentemente detto, attraverso una analisi preliminare dei dati aziendali sono state individuate le regioni, e le relative province, che hanno fatturato di più negli ultimi anni e che hanno mostrato un trend positivo nell'apertura di nuovi punti vendita rispetto alle chiusure. In particolare, le regioni identificate sono Basilicata, Abruzzo e Campania, con un focus sulle province di

Avellino, L'Aquila, Salerno, Potenza e Matera. Dato che queste province hanno dimostrato di essere particolarmente attrattive per l'espansione della GDO, è stato deciso di concentrare tutte le future analisi su di esse.

Per poter identificare tale provincia, è stato necessario utilizzare i diversi indicatori selezionati durante la fase di Data Discovery, in particolare quelli ottenuti tramite la previsione per il 2022. Al fine di condurre un'analisi accurata, è stato deciso di adottare una metodologia di valutazione ponderata.

In primis, dopo uno studio approfondito e una riflessione personale ho stabilito le varie importanze relative da associare a ciascun indicatore , in base alla rilevanza che secondo me ognuno di essi ha nell'ambito dell'analisi.

In base all'ordine di priorità, di seguito viene presentato l'elenco degli indicatori secondo il livello di importanza assegnatogli:

1. **Densità demografica** -> È comunemente considerato uno dei fattori chiave nella valutazione di un'area per l'apertura di un supermercato. Infatti, una maggiore densità demografica implica un numero più elevato di potenziali clienti nelle vicinanze, aumentando in questo modo le opportunità di successo per il nuovo punto vendita. Tuttavia, è importante notare che questo indicatore, se considerato singolarmente, potrebbe non fornire valutazioni ampiamente affidabili, in quanto bisogna prendere in considerazione anche le caratteristiche ,finanziarie e non, della popolazione coinvolta.
2. **Reddito medio delle famiglie** -> In linea a quanto esposto nel punto precedente, il secondo indicatore, in termini di importanza, è rappresentato dal reddito medio. Esso fornisce delle informazioni cruciali sulla situazione economica delle famiglie nelle zone prese in esame. Un reddito più elevato può implicare una maggiore capacità di spesa, risultando quindi un fattore positivo per l'apertura di nuovi supermercati.
3. **Spesa media mensile familiare per beni e servizi alimentari** -> Al fine di acquisire una visione più specifica sulla disponibilità economica, ho ritenuto opportuno considerare la spesa media mensile delle famiglie dedicata ai servizi alimentari. Questo indicatore non solo fornisce informazioni dirette sulla disponibilità finanziaria delle famiglie per l'acquisto di prodotti alimentari, ma anche sulla propensione a dedicare una parte consistente del proprio budget alla spesa alimentare, sottolineando come, indipendentemente dal reddito, le famiglie diano particolare priorità nel soddisfare le necessità alimentari di base e non solo.
4. **Incidenza della povertà** -> Un altro elemento di fondamentale importanza per la valutazione del benessere socioeconomico di un'area è l'incidenza della povertà. Essere situati in una zona ad alto rischio di povertà potrebbe influenzare negativamente le

opportunità economiche delle famiglie residenti. Tuttavia, conoscere quali sono le zone caratterizzate da redditi più bassi e da un indice di povertà più elevato può fornire informazioni di valore per determinare la tipologia di format distributivo più adatta, e di conseguenza quale dei tre brand della GDO è conveniente aprire. Ad esempio, in zone ad alto rischio di povertà risulta particolarmente vantaggiosa l'apertura del brand che offre prezzi più bassi e convenienti, così da soddisfare le esigenze delle famiglie locali.

5. **PIL pro-capite** -> Il PIL pro-capite è un indicatore statistico ottenuto dal rapporto tra il valore del PIL del Paese nel periodo preso in considerazione e il numero di abitanti. Esso fornisce una misura del tenore di vita registrato in media nel Paese.[33] Considerando la sua rilevanza, ho ritenuto tale indicatore significativo per la mia analisi. Un valore più elevato del PIL indica una maggiore capacità delle famiglie di acquistare beni e servizi, inclusi quelli di genere alimentare, e ciò comporta una maggiore opportunità di successo per il nuovo punto di vendita.
6. **Spesa media mensile complessiva per tutti i beni e servizi** -> Oltre alla spesa alimentare ho considerato anche l'indicatore relativo alla spesa media mensile familiare per tutti i beni e i servizi. Nonostante l'apertura di un supermercato si focalizzi principalmente sulla vendita di prodotti alimentari , è comunque relativamente utile tenere in considerazione la spesa totale delle famiglie. Ciò deriva dal fatto che una maggiore propensione all'acquisto generale potrebbe comportare una maggiore propensione all'acquisto di beni alimentari.
7. **Tasso di disoccupazione totale e giovanile** -> Infine, ho posizionato il tasso di disoccupazione (sia giovanile sia totale) all'ultimo posto nella scala dell'importanza relativa. Questa scelta è la dovuta dal fatto che sebbene essi forniscano informazioni sul benessere e la qualità di vita della popolazione, non offrono la stessa precisione e completezza rispetto agli altri indicatori succitati. È bene comunque sottolineare che un tasso di disoccupazione inferiore può indicare una maggiore stabilità economica, un aumento delle opportunità di lavoro e anche una maggiore sicurezza finanziaria per la popolazione, contribuendo così a un aumento della domanda di prodotti alimentari e favorendo la profittabilità del supermercato.

Dopo aver definito l'ordine di importanza degli indicatori come esposto in precedenza, è stato adottato un sistema di assegnazione dei pesi al fine di rappresentare le importanze relative di ciascun indicatore. In particolare, in questo sistema ho attribuito il valore più elevato (7) all'indicatore considerato più importante, quindi la densità demografica, seguito da valori

progressivamente più bassi (6 per il secondo, 5 per il terzo e così via) fino ad arrivare al valore più piccolo (1) assegnato all'indicatore ritenuto meno rilevante, il tasso di disoccupazione.
Prima di calcolare il punteggio pesato per ogni città è necessario fare prima un processo di normalizzazione. Questo processo è fondamentale poiché si sta lavorando con indicatori che presentano unità di misura o scale di valore differenti. Ad esempio, la densità di popolazione potrebbe essere espressa in numero di persone per chilometro quadrato, mentre il reddito potrebbe essere espresso in valuta. Inoltre, quando vengono assegnati dei pesi ai vari indicatori è importante che questi riflettano in modo preciso l'importanza relativa di ciascuno, e, solo grazie alla normalizzazione i pesi assegnati risulteranno applicati in modo equo. Infine, normalizzare i dati rende più semplice e intuitivo il confronto tra vari fattori, permettendo di individuare con facilità i modelli, le tendenze e le possibili relazioni tra di essi, facilitando in questo modo la presa di decisioni basate sui dati.
Nel caso in esame, al fine di metterli tutti su una scala comune è stata utilizzata la normalizzazione min-max, che ridimensiona i valori degli indicatori tra 0 e 1.

$$z = \frac{x - \min(x)}{[\max(x) - \min(x)]}$$

Nella formula descritta z corrisponde al valore normalizzato, x è l'indicatore da normalizzare, min(x) rappresenta il minimo tra tutti i valori dell'indicatore, max(x) rappresenta il valore più grande.
Per il tasso di disoccupazione e per l'incidenza della povertà verrà applicata la formula inversa, poiché contrariamente agli altri indicatori, in questi due casi specifici a valori più elevati è associata una condizione peggiore e viceversa.

$$z = 1 - \frac{x - \min(x)}{[\max(x) - \min(x)]}$$

Dopo aver calcolato il valore normalizzato per ciascun indicatore si può procedere con la valutazione pesata. Moltiplico i valori normalizzati per i rispettivi pesi e successivamente calcolo il punteggio totale per ciascuna città. Per fare questo utilizzo una query che mi permette di sommare i punteggi pesati per ciascuno degli indicatori , consegnandomi una valutazione complessiva della città basata su fattori differenti. Infine, confronto i punteggi totali ottenuti e attraverso la funzione di ranking noto quale è la provincia con il punteggio totale più alto, di conseguenza la migliore, e quella con il punteggio più basso.

RANK	PROVINCIA	PUNTEGGIO
1	Salerno	18,09
2	L'Aquila	15,27
3	Potenza	12,17
4	Avellino	12,14
5	Matera	8,08

Il risultato ottenuto evidenzia quindi come Salerno è la provincia migliore in cui aprire nuovi punti vendita.

Una ulteriore analisi da tenere in considerazione, al fine di ottenere un quadro ancora più completo, è quella relativa ai competitors nella provincia selezionata. Come già visto, durante la fase di Data Discovery sono stati identificati i principali attori della GDO nelle aree prese in esame, in particolare è stato valutato il numero di competitor operanti in ciascuna delle provincie e sono state raccolti i dati relativi alle vie dei competitor, includendo sia le catene nazionali che i negozi indipendenti.

Tali informazioni rivestono un ruolo fondamentale nel determinare quale tra le strategie di posizionamento risulta più vantaggiosa da adottare all'interno della provincia individuata. Il cliente si trova di fronte a due opzioni strategiche differenti, entrambe con i rispettivi vantaggi, svantaggi e relativi rischi.

La prima strategia consiste nell'essere una sorta di "*first mover*" decidendo di posizionarsi in zone del tutto nuove, in cui nessuno dei competitor ha mai aperto punti vendita. La scelta di tale approccio potrebbe comportare un vantaggio unico poiché si entrerebbe in aree senza dover affrontare una concorrenza diretta. In questo modo il cliente potrebbe diventare leader e potrebbe conquistare una fetta di acquirenti fedeli fin da subito. Nonostante ciò, la scelta di tale strategia potrebbe comunque comportare anche dei rischi. Infatti, l'assenza di competitor in quella specifica zona potrebbe essere dovuta da una domanda potenzialmente bassa o poco sviluppata, rischiando quindi un ritorno sugli investimenti più lento. Inoltre, potrebbe essere necessario investire inizialmente grande parte delle risorse in marketing così da poter influenzare maggiormente la scelta dei consumatori e aumentare il valore del nuovo punto vendita.

Il secondo approccio, invece, consiste nell'essere un "*follower"* e quindi la GDO sceglie di posizionarsi nelle zone in cui i competitor hanno già aperto o hanno intenzione di aprire i loro punti di vendita. Il cliente si troverà quindi a competere fin dall'inizio con gli altri player della Grande Distribuzione Organizzata già presenti, al fine di sfruttare un mercato già consolidato e stabile, che presenta una domanda esistente. Il vantaggio che deriva da questa scelta è la certezza di posizionarsi

in zone che presentano già una potenziale clientela e quindi in mercati promettenti e redditizi. Diversi sono gli svantaggi o i cosiddetti rischi che questa scelta può comportare. Innanzitutto, differenziarsi dalla concorrenza potrebbe risultare complicato . In particolare, potrebbe essere difficile accaparrarsi della clientela, già fidelizzata con altri. Bisognerebbe quindi cercare di offrire servizi nuovi e innovativi, prezzi convenienti e offerte accattivanti. Anche in questo caso risulterebbe importante creare una buona campagna di marketing e pubblicità così da farsi conoscere più velocemente.

Come possiamo quindi notare entrambe le strategie presentano opportunità e sfide specifiche. Sicuramente la scelta definitiva dipenderà da una valutazione accurata dei fattori locali, delle risorse disponibili, degli obiettivi aziendali finali e dalla propensione al rischio del nostro cliente.

4. Data Visualization

Con il termine *data visualization* viene indicata la rappresentazione di informazioni e dati mediante l'utilizzo di diagrammi, grafici, mappe e altri strumenti visivi. Negli ultimi anni, questa disciplina è diventata sempre più determinante nell'ambito dell'analisi dei dati, divenendo un pilastro della Business Intelligence.

Nei contesti aziendali la visualizzazione dei dati è un mezzo sempre più importante poiché consente alle persone di comprendere con facilità la grande mole di dati analizzata, risultando una guida indispensabile per il processo decisionale . Grazie ad una rappresentazione visiva risulta più intuitivo identificare le tendenze, i modelli o i valori anomali che si nascondono dietro ogni set di dati. Infatti, alcuni studi dimostrano che l'uomo è in grado di elaborare gli elementi visivi 60.000 volte più velocemente rispetto a quelli di testo. [34]

La visualizzazione dei dati viene utilizzata in innumerevoli campi, dal marketing alle politiche pubbliche, dall'istruzione allo sport. I vantaggi legati dall'utilizzo di tale strumento sono:

- Storytelling: la visualizzazione dei dati permette di raccontare una "storia" attraverso l'utilizzo di colori, pattern, design, grafici. Queste forme visive attirano l'attenzione degli osservatori, suscitando in loro curiosità e maggior interesse in modo da trasmettere efficacemente il messaggio che quelle informazioni vogliono comunicare;
- Accessibilità: la data visualization rende le informazioni accessibili e di facile comprensione per un'ampia gamma di destinatari, anche per chi è meno esperto nel campo specifico. Viene quindi favorita la condivisione e collaborazione tra i team o le diverse parti interessate;
- Visualizzazione delle tendenze: rappresentando le informazioni per mezzo di un grafico o un diagramma risulta più semplice rispetto ad una analisi numerica tradizionale individuare le tendenze e le correlazioni all'interno di un set di dati;
- Esplorazione interattiva: la rappresentazione visiva offre l'opportunità di esplorare in dettaglio i dati. Infatti, grazie all'utilizzo di strumenti interattivi è possibile filtrare, zoomare, scorrere, analizzare i dati da prospettive differenti, favorendo in questo modo una comprensione più approfondita e una collaborazione più efficace [35] ;
- Sintesi delle informazioni: la data visualization consente di rappresentare una grande mole di dati in modo sintetico, semplificando così la comprensione dei vari concetti e permettendo una visione d'insieme;

- Supporto alle decisioni: grazie ad una rappresentazione chiara visiva dei dati è possibile valutare le varie opzioni, confrontare le prestazioni, individuare opportunità oppure indentificare eventuali errori. In questo modo, quindi, la data visualization fornisce una base solida per prendere decisioni informate;

Esistono molti strumenti differenti, adatti ad ogni esigenza, come Tableau, Google Charts, Power BI, Jupyter. Questi strumenti producono un documento che prende il nome di *report.* Esso è una combinazione di tabelle, grafici che presentano le misure rilevanti per i vari fenomeni analizzati e mostrano le informazioni in modo organizzato e comprensibile.

L'implementazione di un processo di reportistica coinvolge diverse fasi, che possono variare in base al contesto o alle esigenze dell'organizzazione. Le fasi principali includono:

- Identificazione delle esigenze informative e di visualizzazione: in questa fase iniziale vengono determinate le varie necessità dell'organizzazione e i vari requisiti di visualizzazione dei dati. Si analizzano le varie tipologie di report, quali sono i dati richiesti e quali possono essere le modalità più appropriate per comunicare le informazioni;
- Identificazione del contesto informativo e delle fonti: vengono esaminate le fonti di dati disponibili e viene valutato il contesto informativo dell'organizzazione. Bisogna determinare quali sono i dati necessari per la creazione dei report, quali sono le fonti affidabili da cui estrarre i dati (possono essere database interni all'azienda o fonti esterne) ;
- Identificazione della configurazione del sistema hardware/software: si deve definire l'infrastruttura tecnologica necessaria per supportare il sistema di reportistica così da pianificare l'implementazione dei vari componenti;
- Fase di integrazione hardware/software delle risorse informative: in questa fase vengono importati i vari dati dalle fonti identificate , vengono elaborati e preparati per la creazione dei report;
- Preparazione del report: viene progettato e sviluppato il layout dei report, determinando quali modelli di visualizzazione (grafici, tabelle, istogrammi, eccetera) saranno inclusi. In questa fase vengono anche definiti i criteri per l'organizzazione e la presentazione dei dati all'interno del report;
- Validazione del report: fase di controllo e test per assicurarsi che i report generati soddisfano le esigenze informative previste. Viene verificata l'accuratezza dei dati e comprensibilità delle visualizzazioni;
- Fase di collaudo del sistema: Si effettuano test approfonditi del sistema di reportistica per verificare la sua stabilità, la sua performance e la sua affidabilità. Vengono simulate diverse situazioni al fine di individuare e risolvere eventuali problemi o errori;

- Fase di esercizio del sistema di reportistica: dopo aver superato le fasi di validazione e collaudo viene messo in funzione il sistema di reportistica;

Per questa tesi è stato deciso di utilizzare Power BI.

4.1 Power BI

Power BI è una suite di strumenti di Business Intelligence sviluppata da Microsoft, ampiamente utilizzata da analisti aziendali e professionisti.

La sua dashboard consente di visualizzare i dati in molti stili differenti, tra cui mappe, grafici, diagrammi, istogrammi, grafici a dispersione e molto altro ancora. La sua interfaccia è progettata in modo semplice ed intuitivo, così da essere facilmente accessibile anche da coloro che non hanno alcuna conoscenza dei dati. Inoltre, sfrutta la funzionalità "AI Insights" al fine di individuare approfondimenti all'interno dei set di dati mediante l'utilizzo dell'intelligenza artificiale.

Power BI facilita la connessione alle diverse origini di dati, consentendo la combinazione in un unico modello dati in modo da creare oggetti visivi di facile interpretazione per chi guarda. Supporta una vasta gamma di fonti, come ad esempio file Excel, file di testo CSV, database, servizi cloud e così via. Inoltre, fornisce anche la possibilità di modificare e trasformare i dati direttamente nella piattaforma, senza intaccare la sorgente originale.

I principali utilizzi di Power BI è possibile riassumerli come segue:

- Creazione di report e dashboard che presentano i set di dati attraverso di elementi visivi;
- Collegamento di varie fonti di dati per creare un unico modello e ottenere approfondimenti aziendali;
- Utilizzo di strumenti avanzati di visualizzazione per trasformare i dati in grafici a barre, grafici a torta, KPI e altri elementi visivi;
- Promozione di una cultura aziendale basata sui dati, svolgendo così un ruolo fondamentale nella presa di decisioni strategiche;
- Collaborazione e condivisione dei report e dashboard con vari membri del team per favorire una comprensione comune e una presa di decisioni condivisa. Le dashboard, infatti, fornisco una panoramica completa delle metriche più importanti per gli utenti aziendali. Essi con un semplice clic riescono ad esplorare cosa si nasconde dietro ogni visualizzazione.

Inoltre, Power BI offre anche la possibilità di accedere ai dati ovunque ci si trovi grazie all'app Power BI Mobile, permettendo così di essere continuamente aggiornati e informati. [36] [37]

4.1.1 Strumenti di Power BI

Power BI può essere definito come un sistema di visualizzazione suddiviso in vari blocchi, che permettono di passare da una analisi generale ad una più dettagliata con un semplice click. Possiamo distinguere principalmente quattro blocchi differenti:

- Visualizzazioni
- Report
- Dashboard
- Dataset

VISUALIZZAZIONI

Come abbiamo già detto per la creazione dei report Power BI offre varie tipologie di oggetti visivi. Le icone di questi oggetti vengono rappresentate all'interno del riquadro *Visualizzazioni*. Inoltre, oltre agli elementi grafici standard, attraverso l'utilizzo del SDK (Software Development Kit, insieme di librerie, strumenti e risorse fornite da Microsoft) gli sviluppatori sono in grado di creare oggetti visivi personalizzati per soddisfare ogni esigenza aziendale. Questi possono essere distribuiti in tre modi differenti:

- File di oggetti visivi personalizzati: gli sviluppatori creano un file contente sia il codice sia le risorse necessarie per ottenere l'oggetto visivo personalizzato. Gli utenti per utilizzare tali elementi dovranno importare su Power BI il file in questione;
- Oggetti visivi di organizzazione: gli sviluppatori possono creare e distribuire le varie visualizzazioni personalizzate all'interno di una organizzazione specifica, in modo che solo gli utenti facenti parte l'organizzazione stessa sono autorizzati ad utilizzarli;
- Oggetti visivi del Marketplace: gli sviluppatori pubblicano e condividono gli oggetti visivi personalizzati all'interno del Marketplace di Power BI, una piattaforma online in cui tutti gli utenti possono scaricarli ed utilizzarli al fine di arricchire le proprie visualizzazioni;

La possibilità di avere visualizzazioni personalizzate fornisce quindi una maggiore flessibilità alle organizzazioni che desiderano comunicare dati specifici o avere determinati requisiti per rappresentare le proprie informazioni.

In sintesi, possiamo dire che la visualizzazione è la rappresentazione visiva dei dati che consente agli utenti di esplorare le informazioni con facilità ed efficacia. Di queste, Power BI ne offre una vasta selezione dalle più semplici come grafici a barre, tabelle pivot, mappe geografiche, alle più avanzate come ad esempio le Heat Map che mostrano la distribuzione dei dati attraverso una scala di colori.

REPORT

Un report è invece una "visualizzazione multi-prospettiva" che riunisce in una o più pagine diversi elementi visivi al fine di rappresentare i risultati e le informazioni di un particolare set di dati . La combinazione di grafici, mappe, metriche e altri oggetti visivi consente di creare una narrazione coerente dei dati e fornisce una visione completa delle informazioni.

Gli oggetti visivi all'interno dei report non sono statici, essi si aggiornano autonomamente in base alle modifiche sottostanti garantendo così informazioni sempre aggiornate e accurate. È inoltre possibile utilizzare la funzione filtro al fine di selezionare esclusivamente le informazioni rilevanti . Questo consente di esplorare e modificare efficacemente i dati presenti nel report senza dover alterare e danneggiare il set di dati originale.

[38]

DASHBOARD

La dashboard è una raccolta di visualizzazioni, chiamate riquadri, che è possibile condividere tra i diversi utenti. A differenza del report, essa è composta da una singola pagina che ospita diverse visualizzazioni.

È progettata per fornire una visione d'insieme delle varie visualizzazioni e metriche chiavi provenienti da un report. Vengono spesso utilizzate per monitorare le prestazioni delle attività aziendali in tempo reale.

È altamente interattiva, permettendo agli utenti di ottenere maggiori dettagli e informazioni approfondite selezionando semplicemente punti specifici nei grafici e interagendo con gli oggetti visivi. Inoltre, le dashboard possono anche essere condivise tra i vari utenti, garantendo così una collaborazione efficace e una visualizzazione delle informazioni sicura e controllata.

[39]

Di seguito possiamo vedere le principali differenze tra i due:

CAPACITA'	DASHBOARD	REPORT
Pagine	Una	Una o più
Origine dei dati	**Uno o più** report o set di dati per dashboard	**Un singolo set** di dati per report
Possibilità di filtrare	No	Si
Possibilità di impostare avvisi	Si, è possibile creare avvisi di posta elettronica da inviare quando vengono soddisfatte determinate condizioni	No
Possibilità di visualizzazione dei campi e delle tabelle del set di dati sottostante	No	Sì, è possibile vedere i campi e le tabelle del set di dati e i valori che si è autorizzati a vedere

DATASET

Un dataset è una raccolta dei dati utilizzata da Power BI per creare delle visualizzazioni, fornendo quindi la base principale per l'analisi. Rappresentano quindi tutti i dati che si nascondano dietri gli oggetti visivi del report. Questi dati non necessariamente provengono da una singola fonte, anzi solitamente è una combinazione di dati di origini differenti. In Power BI è infatti possibile estrarre i dati da qualsiasi sorgente, come ad esempio Excel, Oracle, SQL server, siti web, social network e così via. Una volta importato il dataset Power BI offre inoltre la possibilità di eseguire diverse operazioni di modellazione e trasformazione dei dati, come unire tabelle, definire le relazioni tra tabelle differenti.

4.1.2 Risultati ottenuti tramite Power BI

Vediamo adesso quali tipologie di visualizzazioni sono state utilizzate nella mia analisi.

Al fine di condurre un'analisi più mirata, come descritto nei capitoli precedenti, sono state selezionate tra tutte le province le cinque che hanno mostrato un trend positivo sia in termini di fatturato che di equilibrio tra aperture e chiusure dei punti vendita nel corso degli anni.

Per agevolare questa analisi, è stato utile creare una dashboard che consentisse di visualizzare l'andamento del fatturato delle diverse province.

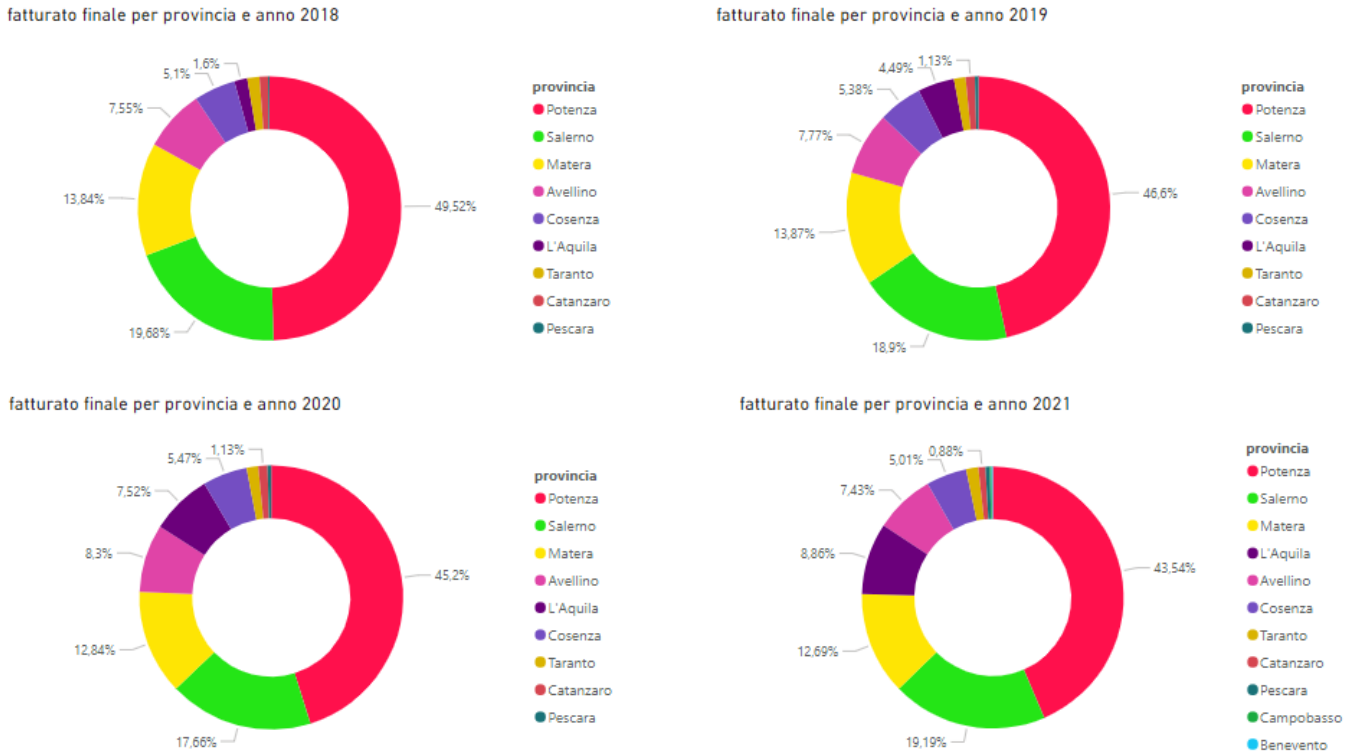

Figura 24: Dashboard fatturato

Da questa analisi è possibile notare che le province di Potenza, Salerno, Matera ed Avellino negli ultimi quattro anni hanno portato significativi fatturati alla GDO. Inoltre, è importante evidenziare come la provincia del L'Aquila negli ultimi anni ha subito un notevole miglioramento, tanto da non solo entrare a far parte delle cinque province di spicco ma anche da riuscire a superare Avellino nell'ultimo anno.

È interessante inoltre vedere come uno stesso set di dati può essere rappresentato tramite diversi oggetti visivi, come possiamo notare nella seguente figura.

provincia	fatturato finale
Avellino	11.255.790,57 €
Benevento	227.374,61 €
Campobasso	312.603,12 €
Catanzaro	1.327.505,85 €
Cosenza	7.591.393,29 €
L'Aquila	13.423.112,84 €
Matera	19.210.737,6 €
Pescara	812.041,87 €
Potenza	65.927.892,1 €
Salerno	29.057.373,0798 €
Taranto	2.279.608,98 €

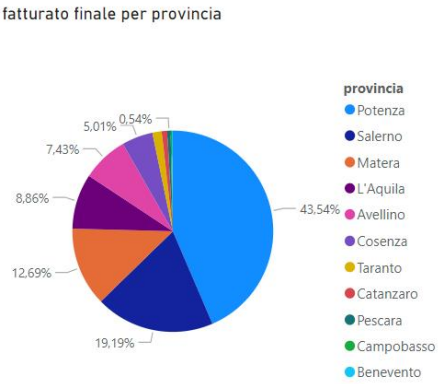

Figura 25: Fatturato rappresentato in diversi oggetti visivi

Per analizzare la distribuzione dei vari supermercati nelle provincie ho utilizzato la **mappa geografica**, un ottimo strumento per poter rappresentare le informazioni geografiche in modo intuitivo ed interattivo. Mediante l'utilizzo delle coordinate geografiche presenti nei dati, Power BI è in grado di posizionare i punti vendita sulla mappa, offrendo una chiara rappresentazione visiva della loro distribuzione geografica. Questo mi ha permesso di individuare facilmente le aree in cui i supermercati sono più concentrati e di confrontarli con i competitor presenti nelle stesse zone.
In un primo momento, ho visualizzato quali tra le diverse provincie possedeva il maggior numero di punti di vendita utilizzando le *bolle* sulla mappa. La dimensione delle bolle rappresenta la quantità dei punti di vendita presenti in una ciascuna provincia, più grande è la bolla maggiore è il numero dei negozi nella provincia corrispondente. Nella mia analisi, la provincia con il maggior numero di punti di vendita è risultata essere Potenza, come evidenziato nella figura sottostante.

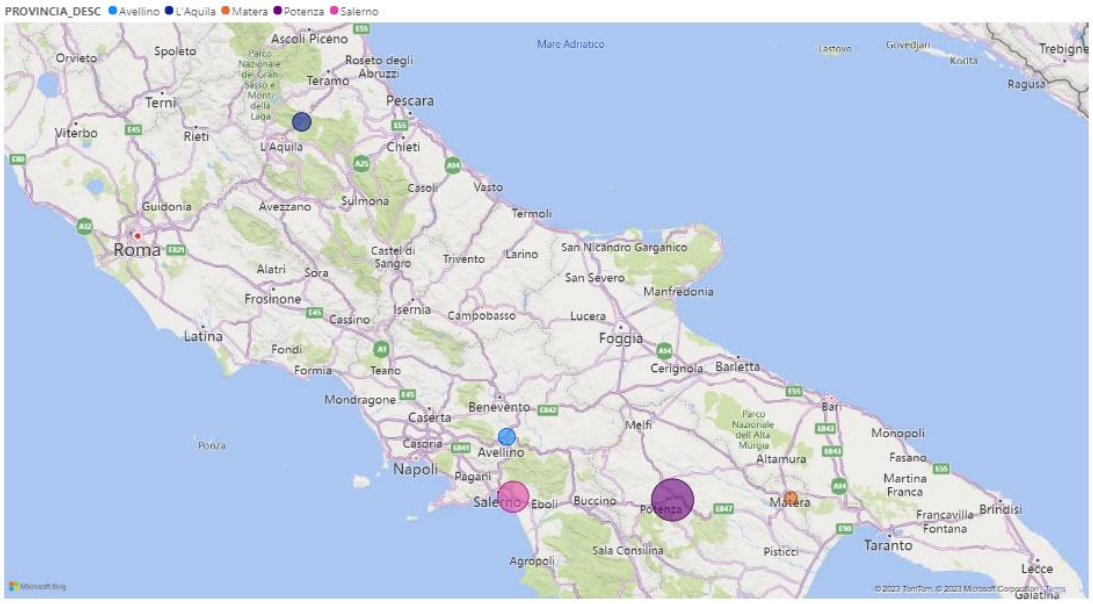

Figura 26: Mappa province

Successivamente ho creato una mappa che mi ha permesso di confrontare la distribuzione dei punti vendita della GDO con i competitor, così da poter comprendere come i diversi supermercati si distribuiscono sul territorio e identificare eventuali opportunità di mercato o discrepanze.

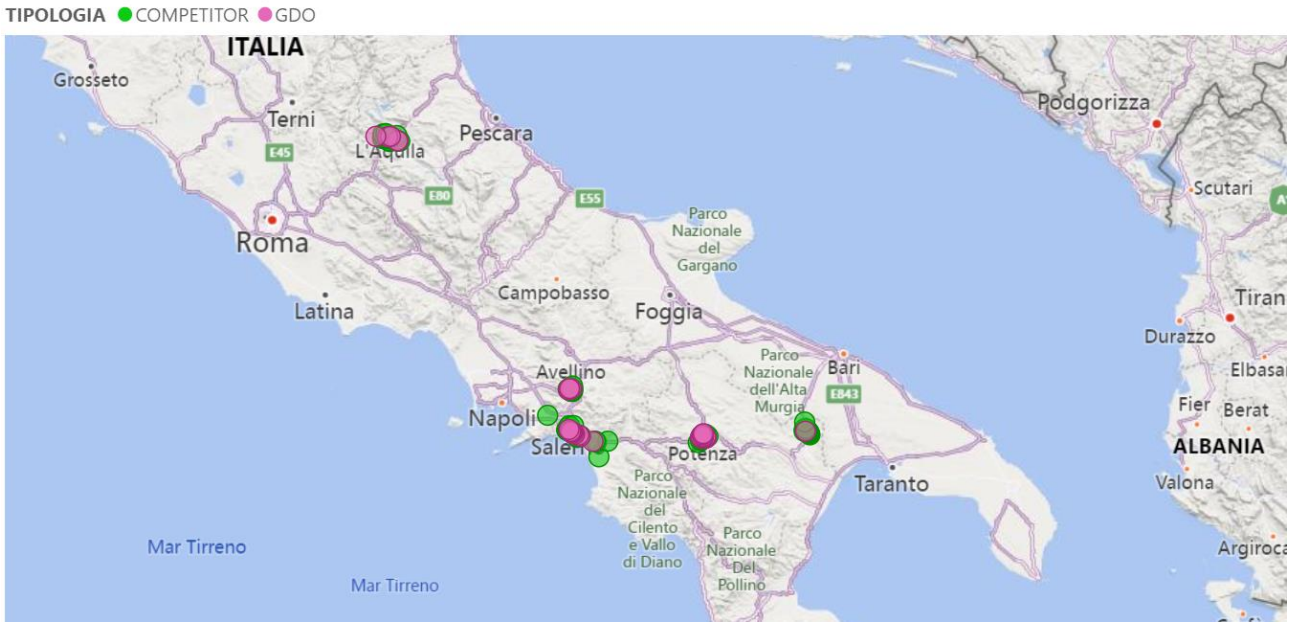

Figura 27: GDO VS competitor

Per analizzare in modo più approfondito ogni singola provincia, ho sfruttato la funzione *filtro* di Power BI. Filtrando per "provincia_desc" , ovvero il nome delle provincie, sono riuscita a visualizzare la distribuzione della GDO e dei competitor provincia per provincia.

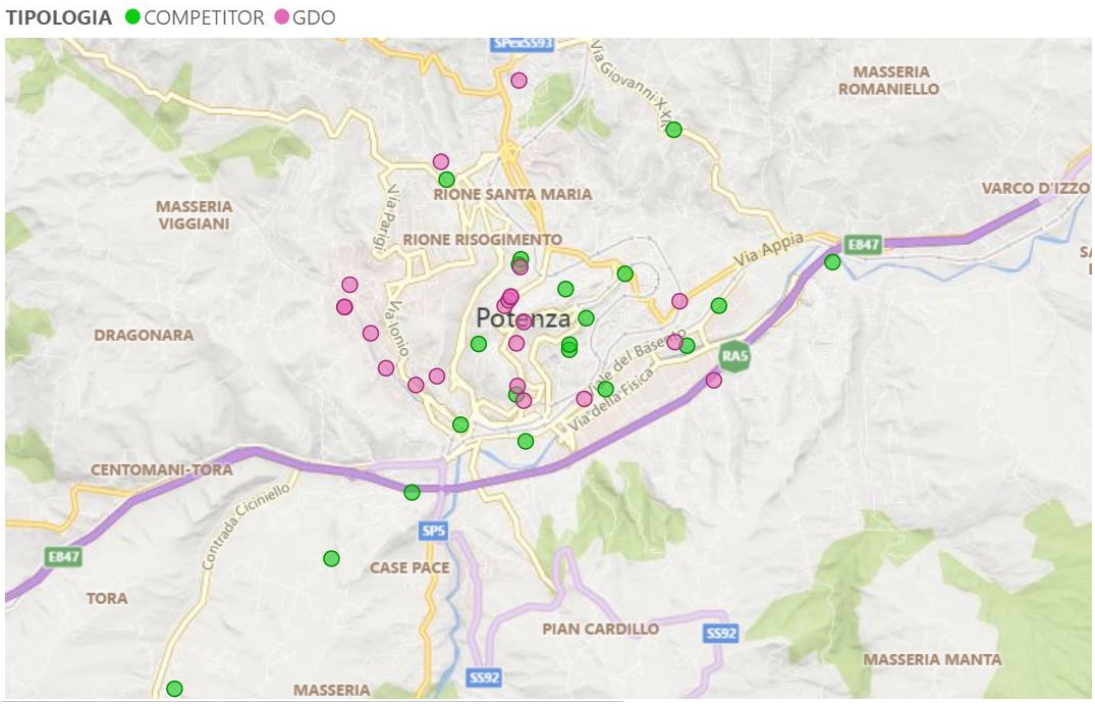

Figura 28: Esempio mappa con dettaglio provincia

Infine, è stata creata un'ulteriore dashboard per visualizzare in modo accurato i cambiamenti previsti nel 2022 degli indicatori, provenienti dalle fonti open.

Come illustrato nei capitoli precedenti, durante il processo di ETL sono state create tre tabelle, una tabella dei fatti e due tabelle delle dimensioni. Per visualizzare graficamente la differenza tra i due anni sono state caricate tutte e tre le tabelle su Power BI, che ha autonomamente riconosciuto la relazione tra la fact table (che contiene tutti gli indicatori e relativi valori per tutte le provincie) e le due tabelle delle dimensioni.

INDICATORE	Anno	Diff %
PIL_PREZZI_CORRENTI	2021	4,10%
SPESA_MEDIA_ALIM_MENSILE_FAM	2021	3,70%
SPESA_MEDIA_TOT_MENSILE_FAM	2021	3,70%
INCIDENZA_POVERTA_RELATIVA	2021	1,30%
TOTALE_PROV	2021	0,00%
TASSO_DISOCCUOAZIONE_GIOV_PROV	2021	-1,33%
REDDITO_MEDIO_FAM	2021	-1,50%
TASSO_DISOCCUOAZIONE_TOT_PROV	2021	-2,71%

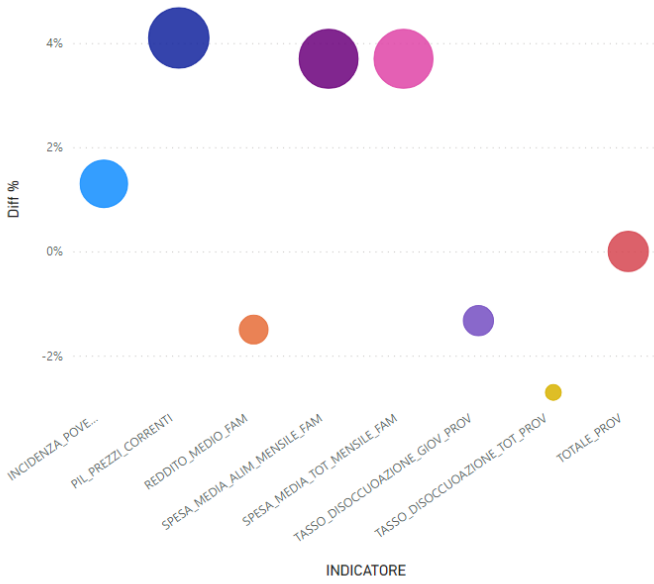

Figura 29:Variazione indicatori

Attraverso il grafico a dispersione è possibile notare facilmente visivamente la variazione prevista per ciascun indicatore. In particolare, la dimensione differente delle bolle evidenzia gli indicatori che hanno subito una variazione maggiore.

Inoltre, è possibile interagire tra i due oggetti visivi attraverso un semplice click. Infatti, se si seleziona l'eventuale indicatore di interesse in uno dei due oggetti , anche gli altri si adatteranno automaticamente mostrando esclusivamente l'indicatore selezionato. Questa interazione tra gli oggetti visivi consente di condurre un'analisi rapida ed efficace. Un esempio è riportato nella figura sottostante, in cui viene selezionata "spesa_media_tot_mensile_fam" nella tabella e il grafico a dispersione si è adattato di conseguenza, fornendo così un'istantanea chiara e focalizzata.

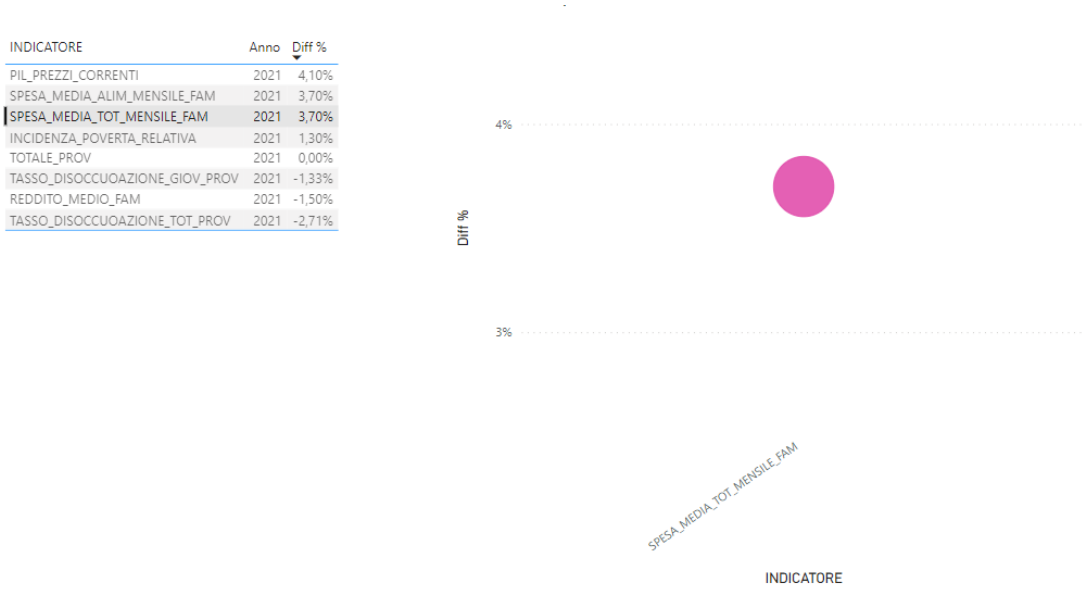

Figura 30: Solo spesa media totale

CONCLUSIONI

Il presente progetto di tesi aveva l'obiettivo di implementare un Data Mart che integrasse i dati sia interni che esterni all'azienda della Grande Distribuzione Organizzata (GDO), al fine di fornire dati puliti e di qualità utili per supportare le future decisioni strategiche dell'azienda.
Nello specifico, attraverso il processo ETL aziendale si voleva creare un Data Mart integrato ideale per la Data Visualization, contenente dati finali di qualità che permettesse di condurre analisi di business significative.
In seguito, con una cross analysis tra i dati open e i dati aziendali interni si aveva l'obiettivo di individuare la posizione ideale per l'apertura dei nuovi punti vendita della GDO.
Dopo le diverse analisi effettuate , è stato inoltre deciso di creare delle Dashboard in Power BI che permettessero di visualizzare in modo intuitivo alcuni dei risultati ottenuti.
Il primo obiettivo di tale progetto è stato raggiunto con successo, poiché il data Mart creato è funzionante e fornisce dati affidabili e utili per le decisioni aziendali. Per quanto riguarda il secondo obiettivo, possiamo affermare di averlo raggiunto in parte. Infatti, grazie all'utilizzo delle query e della previsioni basate su dati open, è stata individuata la provincia strategicamente più conveniente per l'apertura del nuovo punto di vendita, ovvero Salerno. Tuttavia, sono state solo proposte due strategie di posizionamento per scegliere la zona in cui posizionarsi all'interno della provincia stessa, lasciando alla GDO la decisione finale su questa scelta.
In generale, il progetto di tesi ha comunque ottenuto esiti soddisfacenti, fornendo risultati efficaci per le future strategie aziendali. In particolare, questo lavoro mi ha permesso di evidenziare il ruolo fondamentale che i dati giocano nel contesto aziendale odierno. Ho infatti avuto la possibilità di dimostrare come l'utilizzo dei dati aiuta ad ottenere una visione più ampia e completa del contesto operativo aziendale, risultando essenziale per prendere delle decisioni. Il progetto ha dimostrato al cliente come l'analisi dei dati per poter prevedere e decidere dove aprire nuovi punti di vendita forniscano vantaggi significativi alla GDO, come ad esempio la personalizzazione delle offerte, migliori negoziazioni con i fornitori, maggiore fidelizzazione da parte dei clienti.

Bibliografia

Report Doxee. (2020). *La Grande Distribuzione Organizzata: crisi, opportunità e sfide di un settore chiave* .

Adamson, C. (2010). *Star Schema: The Complete Reference.* New York: McGraw-Hill.

Bernice, P. (2013). The emergence of "big data" technology and analytics. *Journal of Technology Research*.

Bregata, L. (2019). *Creation Of A Data Mart For Evaluate The Closing Reasons And The Best Geo-Locations For A Fashion Retail Store, Master's thesis.* Torino: Politecnico di Torino.

Chiarello, D. (2020). *Realizzazione di un Datawarehouse e sviluppo di metodologie di Advanced Analytics a supporto delle strategie aziendali, Master's thesis.* Politecnico di Torino.

Colarieti, M. (2020). *Analisi competitiva tra i players della grande distribuzione organizzata: profili teorici e applicativi, Master's thesis.* Università politecnica delle Marche.

Court, D. (2015). Getting big impact from big data. *McKinsey Quarterly*.

DAMA UK. (2013). DAMA-DMBOK.

DAMA, Global Data Management Community. (s.d.).

D'Ovidio, F., & Villari, M. (2009). *Soluzioni open source per la business intelligence geospaziale.*

Ferrero, G. (2018). *Marketing e creazione del valore* . Torino: Giappichelli Editore.

Fiori, G., & Tiscini, R. (2014). *Economia aziendale* . Milano: Egea.

IBM, c. (2010). *Slide DataStage fundamental.*

Inmon, W. H. (1992). *Building the Data Warehouse.*

Manyika, J., & M, C. (2011). *Big Data: The Next Frontier for Innovation, Competition and Productivity.*

NIELSEN. (2008). *Guida indicatori retail* . The Nielsen Company.

Noce, L., & D'Ercole. (2000). *Data Warehousing.*

Panza, R. (2013). *Manuale di progettazione per la grande distribuzione. Strategie, immagine e format per nuovi consumatori.* Milano: FrancoAngeli.

Porter. (1985). *Competitive Advantage: Creating and Sustaining Superior Performance* .

Raghupathi, W., & V., R. (2014). *Big data analytics in healthcare: promise and potential. Health Information Science and Systems.*

Rezzani. (2012). *Business Intelligence- processi metodi, utilizzo in azienda.* milano: Feltrinelli editore.

Saul, J., Alan, D. ,., Melody, C., & Ted, F. (2020). *Gartner: "5 Steps to Build a Business Case for Continuous Data Quality Assurance".*

Sbrana, R., & Gandolfo, A. (2007). *Contemporary retailing: il governo dell'impresa commerciale moderna.* Torino: Giappichelli editore.

Thompson. (2009).

Tieri, E., & Gamba, A. (2009). *La grande distribuzione organizzata in Italia.* Funzione studi del Banco Popolare.

Sitografia

[1]https://www.allianz-trade.com/en_global/news-insights/business-tips-and-trade-advice/advantages-and-disadvantages-of-expanding-a-business.html

[2] https://fmcggurus.com/blog/fmcg-gurus-the-evolution-of-e-commerce-in-2022/

[3]https://www.peppercontent.io/blog/fmcg-digital-marketing-tips/

[4] https://www.shopify.com/it/blog/analisi-swot

[5] https://asana.com/it/resources/swot-analysis

[6] https://www.bhype.it/il-valore-aggiunto-di-una-indagine-di-mercato-per-aprire-un-negozio/

[7]https://www.almalaboris.com/organismo/blog-lavoro-alma-laboris/67-export-management/1808-ricerche-di-mercato-cosa-sono-come-si-fanno.html

[8] https://www.qualtrics.com/it/experience-management/ricerca/ricerca-di-mercato-una-guida/

[9] https://www.bhype.it/il-valore-aggiunto-di-una-indagine-di-mercato-per-aprire-un-negozio/

[10]https://www.slideshare.net/beppepol/geomarketing-la-dimensione-spaziale-che-migliora-le-vendite-il-marketing-la-strategia

[11]https://www.flyip.it/le-5v-dei-big-data-le-caratteristiche-di-una-massa-di-dati/

[12] https://www.oreilly.com/library/view/the-enterprise-big/9781491931547/ch01.html

[13] https://www.oracle.com/it/database/what-is-a-data-warehouse/

[14] https://www.oracle.com/it/database/what-is-a-data-warehouse/

[15] https://www.ibm.com/it-it/topics/data-warehouse

[16] https://it.wikipedia.org/wiki/Schema_a_stella

[17] https://it.wikipedia.org/wiki/Schema_a_fiocco_di_neve

[18] http://www-db.disi.unibo.it/courses/SIG/DW1.pdf

[19] https://www.talend.com/it/resources/what-is-etl/

[20] https://www.oracle.com/it/integration/what-is-etl/#difference

[21]https://azure.microsoft.com/it-it/resources/cloud-computing-dictionary/what-is-data-integration/

[22] https://www.informatica.com/hk/data-integration-magic-quadrant.html

[23]https://www.ibm.com/it-it/products/datastage

[24]https://www.ibm.com/it-it/topics/data-mart

[25]https://www.bucap.it/news/approfondimenti-tematici/gestione-del-magazzino/top-down-e-bottom-up-i-due-approcci-alla-progettazione-del-data-warehouse.htm

[26]https://vitolavecchia.altervista.org/che-cosa-sono-le-proprieta-acid-dei-dbms-in-informatica/

[27]https://vitolavecchia.altervista.org/caratteristiche-e-utilizzo-di-un-modello-multidimensionale-o-dfm-in-informatica/

[28] https://www.bnova.it/business-intelligence/data-ingestion-significato-esempi-e-vantaggi/

[29]https://www.irion-edm.com/it/data-management/metadata-metadati-che-cosa-sono/
[30]https://www.irion-edm.com/it/data-management/data-quality-che-cose-perche-adottarla-come-applicarla/
[31] https://learn.microsoft.com/it-it/power-bi/guidance/star-schema
[32] https://www.techtarget.com/searchdatamanagement/definition/primary-key
[33]https://www.rgs.mef.gov.it/VERSIONE-I/e_government/amministrazioni_pubbliche/igrue/PilloleInformative/economia_e_finanza/
[34] https://www.studiosamo.it/12-statistiche-conoscere-visual-content-marketing/
[35] https://www.coursera.org/articles/data-visualization
[36] https://www.coursera.org/articles/what-is-power-bi
[37] https://learn.microsoft.com/it-it/power-bi/fundamentals/power-bi-overview
[38] https://learn.microsoft.com/it-it/power-bi/consumer/end-user-reports
[39] https://learn.microsoft.com/it-it/power-bi/consumer/end-user-dashboards
[40] https://esploradati.istat.it/databrowser/#/it/dw/categories
[41] https://www.mef.gov.it/index.html
[42] https://www.bancaditalia.it/pubblicazioni/economie-regionali/
[43] https://www.statista.com/