POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Progettazione e sviluppo di un acceleratore data analytics per l'ambito retail

Relatore
Prof. Cerquitelli Tania

Candidato
Di Varano Kevin

Anno Accademico 2023-2024

Alla mia Famiglia
A tutti i miei Amici

Sommario

L'obiettivo di questa tesi è progettare e sviluppare un acceleratore data analytics per l'ambito retail basato su tecnologie cloud. L'acceleratore è un insieme di strumenti e tecnologie ideate per velocizzare e ottimizzare il processo di analisi dei dati. Tali strumenti possono includere software specializzati, algoritmi di elaborazione dati avanzati, infrastrutture di calcolo ad alte prestazioni e best practices per la gestione dei dati.

L'acceleratore permette alle aziende di individuare più rapidamente tendenze, modelli e insight significativi e, inoltre, consente agli analisti di concentrarsi su attività più strategiche attraverso l'automatizzazione di alcune fasi del processo. Dunque, l'acceleratore considera i punti in comune delle diverse progettualità e standardizza un approccio che può essere adattato in modo flessibile a differenti realtà o casi d'uso.

L'architettura dell'acceleratore ha inizio con l'ETL (Extract-Transform-Load), un processo che prevede l'estrazione, la trasformazione e il caricamento dei dati in un data warehouse, repository contenente dati relazionali in formato tabellare. L'architettura si conclude con l'implementazione della UCV (Unique Customer View), vista unica del cliente che consente di visionare e gestire tutti i dati inerenti ai clienti in un luogo univoco centralizzato.

L'argomento di fondo su cui verterà principalmente la tesi sarà la Customer Analytics, analisi focalizzata ed incentrata sul cliente che viene supportata mediante l'utilizzo dei KPI (Key Performance Indicators). L'ambito di riferimento è il Retail, termine che indica tutte quelle attività associate alla vendita di prodotti/servizi da parte di un'azienda (detta retailer) direttamente al consumatore finale. Nella definizione rientrano soprattutto la Grande Distribuzione Organizzata o GDO (ovvero i supermercati), la Grande Distribuzione Specializzata o GDS (che a differenza delle GDO operano in un unico settore), il fashion e il beauty.

Ringraziamenti

Desidero esprimere la mia profonda gratitudine a tutti coloro che mi hanno accompagnato durante il percorso di realizzazione di questa tesi.

Un ringraziamento speciale va alla mia famiglia e ai miei compagni universitari per il completo sostegno durante questi anni di studio, senza il quale non sarei riuscito a raggiungere questo traguardo.

Vorrei inoltre ringraziare l'azienda Horsa Insight per avermi concesso l'opportunità di condurre la mia tesi presso di loro e il mio relatore, la Professoressa Tania Cerquitelli per i suggerimenti e le preziose indicazioni che mi hanno guidato durante questi mesi di ricerca.

Infine, desidero ringraziare i miei amici, che hanno condiviso con me gioie e sfide durante il percorso di studio: la loro presenza, affetto e incoraggiamento hanno reso questo cammino più piacevole e stimolante.

Indice

Elenco delle figure

Capitolo 1

Introduzione

In questi ultimi anni, la diffusione dei Big Data ha preso piede in diversi settori cambiando in modo significativo il modo di lavorare di molte aziende, che devono gestire la rapidità con cui questi enormi volumi di dati vengono generati attraverso molteplici fonti di diversa natura.

L'obiettivo di questa tesi è quello di andare a descrivere le tecnologie di Data Platform e in particolare il modo in cui esse possono aiutare le aziende del mondo Retail a raccogliere ed analizzare le informazioni che le caratterizzano. Durante i vari capitoli saranno affrontate diverse tematiche.

Nel secondo capitolo, si partirà dalla definizione dei Big Data e di una tecnologia che permette di gestirli, appunto la Data Platform. Successivamente, verranno introdotti i concetti di Data Lake e Data Warehouse con relative similitudini e differenze. In seguito, verrà confrontata l'elaborazione batch con l'elaborazione in real time e infine, verrà descritto il settore della Business Intelligence e tutti i principali software che esso può offrire.

Nel terzo capitolo, si andrà a spiegare il motivo per il quale il cloud è un argomento sempre più d'attualità ed in particolare come esso aiuta a livello di tecnologia. Inoltre, verranno analizzati gli obiettivi, i funzionamenti e i vantaggi che vengono portati avanti da temi sempre più rilevanti nella società moderna, ovvero l'Intelligenza Artificiale, il Machine Learning e l'Intelligenza Artificiale Generativa.

Nel quarto capitolo, si andrà a parlare di un concetto peculiare nel mondo Retail (ambito su cui si concentrerà la tesi) ossia quello relativo alla Customer Analytics. Quest'ultimo aspetto risulta più interessante rispetto ad altre tematiche comuni e ricorrenti; si pensi ad esempio all'analisi delle vendite, concetto più trasversale e che quindi viene applicato anche in altri settori.

Nello specifico, in questa sezione si affronteranno diversi temi tra cui la definizione della UCV (Unique Customer View) e relativi KPI, la targetizzazione dei clienti a supporto di una campagna marketing, la personalizzazione dei messaggi di comunicazione (volantino personalizzato), il Churn Rate (tasso di abbandono del cliente), il CLV (Customer Lifetime Value) e il TtNP (Time to Next Purchase).

Nel quinto capitolo, verrà illustrata l'architettura dell'acceleratore data analytics basata su tecnologie cloud e tutte le fasi (ingestion, storage, trasformazione, creazione data warehouse, implementazione UCV) che vanno a comporre l'intero processo a cui è sottoposto il dato. Il tutto termina con la fase di visualizzazione, che viene portata avanti attraverso le dashboard, strumenti visivi che si focalizzano sull'evidenziare i dati mediante i KPI più importanti inseriti all'interno della vista unica del cliente.

Nel sesto capitolo, verranno analizzate due applicazioni di casi d'uso in due contesti differenti. Nel primo scenario verrà integrata all'acceleratore la tecnologia Google mentre nel secondo scenario verrà impiegata la tecnologia AWS (Amazon Web Services). Entrambe le tecnologie, insieme a Microsoft Azure, rientrano attualmente tra i migliori cloud provider sul mercato.

L'obiettivo di quest'ultimo capitolo (ma un po' di tutta la tesi) è dunque quello di mostrare come l'acceleratore possa adattarsi a casi d'uso differenti sia a livello di tipologia di cliente coinvolta (contesto GDO, GDS, fashion, beauty) sia appunto a livello di tecnologia impiegata.

Capitolo 2

Data Platform

2.1 Big Data

Nel linguaggio comune quando si sente parlare di Big Data si pensa immediatamente a grossi e complessi volumi di dati provenienti da svariate fonti, la cui gestione da parte dei software di elaborazione dati tradizionale non sussiste più. Per completare la definizione di Big Data è necessario introdurre il concetto meglio conosciuto come le 3 V (o 5 V).

Figura 2.1: Big Data
(20)

2.1.1 Le 3 (+2) V dei Big Data

Ecco l'elenco che descrive le 3 V (recentemente ne sono state aggiunte due quindi in realtà possono essere considerate 5):

1. Volume
   Con i Big Data bisogna elaborare grosse moli di dati che sono raccolte da sorgenti differenti tra cui ad esempio transazioni commerciali, transazioni bancarie, movimenti sui mercati finanziari, dispositivi intelligenti (IoT), social media, click sul web, video e così via.

   Ad oggi, l'archiviazione (ed il relativo costo) è indubbiamente più gestibile grazie a piattaforme come i data lakes. A livello di numeri si varca la soglia delle centinaia di terabyte (1 TB = 1000 GB) o petabyte (1 PT = 1000 TB);

2. Velocità
   La sfida delle aziende è quella di essere in grado di collezionare questi dati in maniera del tutto rapida e di conseguenza valutarli ed analizzarli nel più breve tempo possibile (in tempo reale), affinchè possano essere prese decisioni di business in maniera tempestiva.

   La velocità più elevata dei dati è indirizzata direttamente nella memoria invece di essere scritta su disco e per gestire questa rapidità di immagazzinamento dati in maniera corretta c'è bisogno di strumenti ad hoc;

3. Varietà
   La varietà si riferisce alle diverse tipologie di dati messi a disposizione, ad esempio sistemi transazionali e gestionali aziendali, social network, siti web, open data e così via. Alcuni di questi sono dati strutturati e si adattano in maniera impeccabile ad un database relazionale.

   Con l'avvento dei Big Data, i dati arrivano anche semistrutturati o non strutturati (principalmente testo, audio e video). Per estrarre valore da questa tipologia di dati, è necessaria un'elaborazione preliminare addizionale affinchè si possano generare degli insight rilevanti.

   Tuttavia, per molti esperti del settore è necessario considerare due dimensioni ulteriori:

4. Variabilità
   Oltre alle caratteristiche sopra menzionate, i flussi di dati sono imprevedibili perciò variano di frequente e in modo costante. C'è bisogno di saper contestualizzare i dati poichè l'interpretazione di un dato può cambiare a seconda della situazione in cui viene raccolto e analizzato.

   La variabilità si manifesta anche a seconda del momento in cui viene messa in atto l'analisi, la quale pertanto deve essere svolta real time;

5. Veridicità
   La veridicità si riferisce all'attendibilità, affidabilità e qualità dei dati. Poiché i dati provengono da sorgenti disparate, è una missione ardua mettere in relazione, pulire e trasformare i dati tra i diversi sistemi. Oltretutto, è facile incorrere in situazioni che vanno ad inficiare sulla veridicità del dato.

   Ad esempio si possono eseguire dei click per errore o possono venire usati gli stessi dispositivi da persone differenti. Bisogna sempre valutare con molta precisione e attenzione i Big Data, analizzarli scrupolosamente e confermarne la veridicità.

2.1.2 Casi d'uso dei Big Data

I Big Data possono dare supporto nell'affrontare una serie di attività aziendali, dalla customer experience agli analytics. Eccone alcune (4):

- Sviluppo del prodotto
  Diverse organizzazioni sfruttano i Big Data per anticipare la domanda dei clienti. Per far questo vengono realizzati modelli predittivi per nuovi prodotti e servizi, classificando attributi chiave del passato con quelli attuali e modellando la relazione tra tali attributi e il successo commerciale delle offerte;

- Customer Experience
  I Big Data ti permettono di raccogliere dati da social media, siti web, registri e altre fonti per ottimizzare l'esperienza di interazione e massimizzare il valore fornito. In questo modo, il tasso di abbandono dei clienti calerà drasticamente e i problemi verranno gestiti in maniera non passiva assegnando pertanto offerte personalizzate ai fini della fidelizzazione;

- Frode e compliance
  I Big Data supportano l'individuazione dei modelli dati che indicano frodi e l'aggregazione di grandi volumi di informazioni per rendere i rapporti normativi decisamente più tempestivi. Tuttavia, nel momento in cui si parla di sicurezza, si fa riferimento ad intere squadre di esperti hacker perciò tali scenari più i requisiti di compliance (conformità a determinate norme, regole o standard) sono in costante evoluzione;

- Machine Learning
  Oggigiorno è possibile insegnare alle macchine piuttosto che limitarsi a programmarle. Questo è il compito che spetta al Machine Learning, il quale attraverso determinati modelli è in grado di gestire con relativa facilità la grossa mole derivante dai Big Data;

- Efficienza operativa
  Non è di per sè un argomento di innovazione ma è un'area in cui i Big Data stanno avendo maggior influenza ed effetto. I Big Data possono essere impiegati anche per ottimizzare il processo decisionale in linea con l'attuale domanda di mercato, analizzando ed esaminando la produzione, il feedback e altri fattori che riguardano il cliente per limitare le interruzioni e anticipare le richieste future.

2.1.3 Funzionamento dei Big Data

I Big Data procurano nuovi insight che portano a nuove occasioni e modelli di business. Per iniziare sono necessarie le seguenti azioni (4) - (5):

1. Integrare
   Come già ribadito più volte in precedenza, le sorgenti dei Big Data sono tra le più svariate. Durante l'integrazione bisogna inserire i dati, elaborarli e verificare che siano formattati e messi a disposizione in una forma con cui gli analisti aziendali possano lavorare.

   Di conseguenza, i tradizionali meccanismi di integrazione (tra cui il processo ETL che sta per estrazione, trasformazione e caricamento dei dati) spesso non sono all'altezza del compito. Sono richieste nuove strategie e tecnologie per

analizzare i set di Big Data su scala terabyte (1 TB = 1000 GB) o petabyte (1 PT = 1000 TB);

2. Gestire
   I Big Data esigono spazio di archiviazione. Molte persone scelgono la loro soluzione di archiviazione in base a dove risiedono attualmente i loro dati.

   Essa può essere on-prem ma ultimamente il cloud sta ottenendo molta fama poichè supporta i requisiti di calcolo correnti e permette di incrementare le risorse secondo necessità. In alcuni casi potrebbero essere sfruttate entrambe le soluzioni;

3. Analizzare
   Con tecnologie ad alte performance, le aziende possono valutare di sottoporre tutti i loro Big Data all'analisi. Tuttavia, potrebbe essere più corretto individuare anticipatamente quali dati siano effettivamente utili prima che essi vengano analizzati.

   In ogni caso, l'analisi dei Big Data è il modo in cui le aziende estraggono significato e insight dai dati in formato grezzo. Sempre più frequentemente, i Big Data promuovono le attuali evoluzioni delle advanced analytics, come ad esempio l'intelligenza artificiale;

4. Prendere decisioni migliori e data-driven
   Nel momento in cui i dati sono ben gestiti e affidabili, di conseguenza possono e devono essere prodotte analisi e decisioni che risultano anch'esse affidabili. L'approccio seguito dalle aziende affinchè possano rimanere competitive è quello di operare in maniera data driven.

   Ciò vuol dire che esse devono prendere decisioni di business basate sulle prove concrete proposte dai Big Data piuttosto che basate sulla soggettività. Solo in questa maniera le stesse aziende saranno in grado di estrarre valore significativo dai Big Data, ragion per cui si manifesterà una maggior redditività ed efficacia dal punto di vista operativo.

Nel caso dei Big Data si distinguono 4 classi di Analytics, che applicano strumenti e metodi di analisi differenti (81):

- Descriptive Analytics: strumenti per descrivere la situazione attuale e passata dei propri processi aziendali;
- Predictive Analytics: strumenti per rispondere a domande relative a ciò che potrebbe accadere in futuro;
- Prescriptive Analytics: modelli di ottimizzazione in grado di prevedere degli scenari futuri, in base alle analisi svolte;
- Automated Analytics: strumenti in grado di implementare autonomamente l'azione proposta, in base al risultato delle analisi svolte.

2.2 Definizione Data Platform

2.2.1 Che cos'è una Data Platform

Una data platform è una piattaforma tecnologica integrata finalizzata alla raccolta, elaborazione ed analisi dei dati, che assicura la coesistenza e l'interoperabilità di risorse di dati on-premise (installati e gestiti attraverso computer locali) oppure ospitati nel private/public cloud. Ovunque siano allocati, i dati sono disponibili per gli utenti che potranno in tal modo sfruttare il beneficio derivante da questo mix di risorse, individuando di volta in volta quelle più adeguate per esaudire particolari esigenze. (61)

Tale struttura consente alla piattaforma l'accesso ai dati provenienti da più fonti ed inoltre evita inutili ritardi legati alla loro elaborazione. In aggiunta, è presente un cospicuo livello di sicurezza derivante dal beneficio di poter tenere sempre d'occhio i diversi accessi eseguiti.

Per un processo decisionale più efficente e cooperativo (vista la possibilità di poter condividere dati interdipartimentali), è possibile ipotizzare dashboard, report e avvisi che assistono l'uso e la consultazione dei dati. La data platform assicura una certa agilità, constatabile non solo dall'agevolazione della gestione dei fornitori ma soprattutto dal perfezionamento della data governance (tutto ciò che comprende le azioni da intraprendere, i processi da seguire e la tecnologia di supporto durante l'intero ciclo di vita dei dati).

Fino a poco tempo fa, la maggior parte delle organizzazioni conviveva in silos di dati non scalabili, duplicati e non aggiornati oltre che bloccati in soluzioni proprietarie con un unico livello di sicurezza.

Con le relative piattaforme, il mondo Big Data si pone l'obiettivo di venire a capo di questo problema sfruttando una combinazione di tecnologie scalabili e interoperabili che collaborano simultaneamente per rispondere alle diverse necessità di un'azienda.

2.2.2 Tipologie di Data Platform

Esistono differenti tipologie di data platform, alcune delle quali simili tra loro (57):

1. Enterprise Data Platform (EDP): fornisce un accesso centralizzato alle risorse dati di un'azienda. Le fonti dati sono tradizionali e gestite sia on-premise che in cloud;

2. Modern Data Platform: è l'evoluzione naturale della precedente, in quanto possiede funzionalità più flessibili e idonee al futuro. Si trovano anche soluzioni on-premise, ma più frequentemente vengono impiegate tecnologie cloud (più accessibili a livello di costi);

3. Cloud Data Platform: è un termine ad alto livello per segnalare piattaforme di analisi basate esclusivamente sul cloud computing;

4. Costumer Data Platform: sono piattaforme per l'analisi di dati interamente focalizzate sul cliente. Esse sfruttano fonti eterogenee di dati come ad esempio il Customer Relationship Management (abbreviato con l'acronimo CRM), il quale è utile per la gestione di tutti i rapporti e le interazioni di un'azienda che hanno luogo con i clienti potenziali ed esistenti.

   Altre tipologie di fonti dati sono i social media, i siti web, i sistemi transazionali e l'e-commerce. A seconda del contesto, è possibile descrivere il cliente nei diversi comportamenti attuati grazie all'aggregazione dei dati che crea una profilazione dell'individuo a tutto tondo;

5. Big Data (Analytics) Platform: grazie ad un insieme di features e servizi permette query complesse e grossi volumi di informazioni. È un ambiente

in fase di diffusione e affermazione che sfrutta la coesistenza di differenti tecnologie in un'unica architettura scalabile, sicura e disponibile.

Questa piattaforma si basa anche sul concetto di automazione dell'infrastruttura (in breve tempo creazione da zero di tutti gli elementi fisici) e dell'onboarding dei dati (creazione di pipeline a partire da modelli standard).

2.3 Data Lake

2.3.1 Definizione di Data Lake

Un data lake è un repository con il compito di ingestionare enormi set di dati non elaborati di varia tipologia nel loro formato grezzo. I data lake offrono una strategia di gestione dei dati che torna sempre più comoda alle aziende che desiderano adottare un repository globale ad elevata capacità. (56)

Per dati non elaborati si intendono quei dati che non sono stati ancora trasformati per un particolare obiettivo. In un data lake, un dato non viene definito fino al momento in cui non viene eseguita una query che lo coinvolga. I data scientist possono aver accesso ai dati non elaborati impiegando strumenti di modellazione predittiva e di analisi avanzata.

Nel caso dei data lake non esistono restrizioni o vincoli temporali per l'analisi dei dati, a differenza di quando i dati vengono elaborati a seconda del loro particolare obiettivo. Di conseguenza, a nessun dato viene applicato un filtro o una rimozione prima di eseguire lo storage nel data lake.

Solo nel momento in cui occorre analizzare un dato è necessaria la sua trasformazione ed elaborazione. È in seconda battuta che viene applicato ad esso uno schema per procedere alla relativa analisi. Tale schema è definito "schema on read", poichè i dati vengono elaborati esclusivamente nel momento in cui sono a disposizione per essere impiegati.

I data lake permettono agli utenti di accedere e analizzare i dati dove e come sono, senza la necessità di trasferirli in un altro sistema. Le informazioni ricavate dai data lake vengono elaborate a seconda dei casi e non estratte in maniera regolare da un'altra piattaforma/repository di dati.

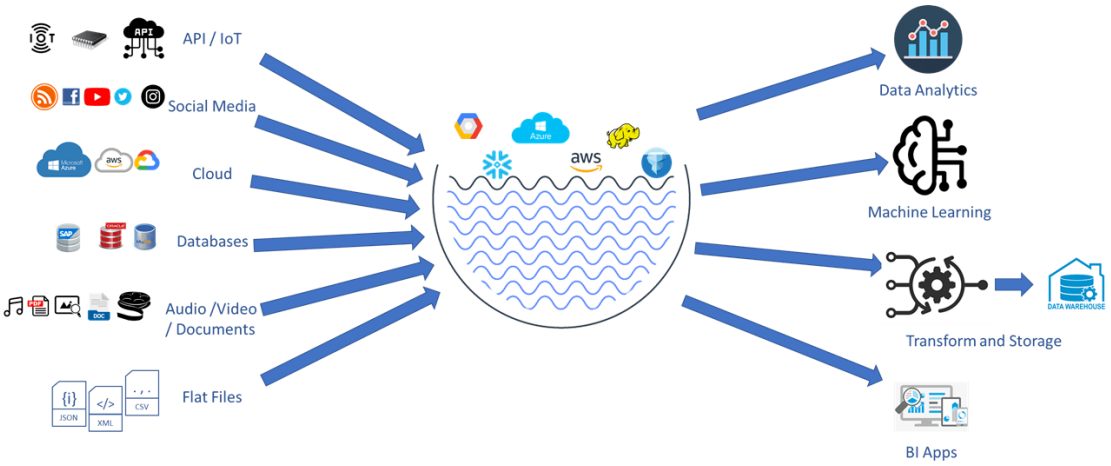


Figura 2.2: Data Lake
(27)

Inoltre, se c'è l'esigenza di dover duplicare un report, gli utenti sono in grado di applicare uno schema e una procedura di automazione. È richiesta non solo una governance ma anche una costante manutenzione dei data lake, altrimenti i dati rischierebbero di diventare inaccessibili, invadenti e poco economici piuttosto che accessibili e utilizzabili. In tal caso si parla di data swamp ovvero palude di dati.

Un data lake fornisce una piattaforma scalabile e affidabile che permette alle aziende di (16):

- Importare dati a qualsiasi velocità da qualsiasi sistema, a prescindere dal fatto che i dati provengano o meno da sistemi on-premise, cloud o edge-computing (l'elaborazione dei dati avviene il più vicino possibile a dove i dati vengono generati, migliorando i tempi di risposta e risparmiando sulla larghezza di banda);

- Archiviare con precisione qualsiasi tipo o volume di dati;

- Elaborare i dati in real time o in modalità batch (sistema per eseguire elevati volumi di job sui dati di tipo ripetitivo, che consente di elaborare i dati

attraverso un numero sufficiente di risorse e con un'interazione minima o nulla da parte dell'utente);

- Analizzare i dati utilizzando SQL, Python, R o qualsiasi altro linguaggio/applicazione per poter fare analisi.

Se l'obiettivo di un'azienda è sfruttare un contesto che sia il più vasto possibile, allora l'utilizzo di un data lake è la scelta ottimale in quanto quest'ultimo non si limita ad archiviare dati ad alta precisione, ma si impegna anche nel procurare agli utenti informazioni più dettagliate sulle situazioni aziendali, permettendo loro di velocizzare gli esperimenti di analisi.

Le aziende si affidano ai data lake per contribuire inoltre a:

- Abbassare il costo totale di proprietà;
- Semplificare la gestione dei dati;
- Prepararsi a incorporare l'intelligenza artificiale e il machine learning;
- Migliorare la sicurezza e la governance.

2.3.2 Architettura di un Data Lake

Come già detto in precedenza, un data lake può essere on-premise o su cloud. I dati possono essere strutturati, semi-strutturati o non strutturati e vengono prelevati da sorgenti differenti all'interno di un'architettura del tutto piatta.

Per la natura di quest'ultima, i data lake consentono una scalabilità massimale fino alla scala exabyte (10^{18} byte, ovvero un trilione di byte). Ciò è fondamentale poichè nel momento in cui si crea un data lake non si sa anticipatamente la mole di dati che verrà ingestionata. Questo tipo di scalabilità non è permessa nei sistemi di storage di dati tradizionali.

Chi trae beneficio da questa tipologia di architettura è la categoria dei data scientist poiché, grazie all'impiego di strumenti di analisi dei Big Data e di machine learning, è permesso loro l'accesso ai dati dell'intera azienda. Inoltre, queste figure professionali possono fare riferimenti incrociati (anche tra dati eterogenei da campi differenti) analizzando e condividendo tutto il necessario per procurarsi nuovi insight.

La governance dei dati rimane un altro tassello essenziale, nonostante non venga a loro applicata una struttura predefinita prima di essere raccolti in un data lake. Una volta inseriti nei data lake, i dati devono essere contrassegnati con metadati in modo tale da assicurare la loro futura accessibilità evitando il fenomeno della data swamp (risultato di un data lake mal gestito, ovvero che manca di adeguate pratiche di qualità e governance dei dati per fornire informazioni approfondite).

2.4 Data Warehouse

2.4.1 Definizione di Data Warehouse

Come sostegno alle attività di business intelligence (BI), un data warehouse è un repository contenente enormi volumi di dati storici centralizzati e consolidati, progettato unicamente per eseguire interrogazioni (query) e analisi attraverso un modello pensato ed organizzato per un preciso scopo. I dati all'interno di un data warehouse provengono da diverse fonti come ad esempio le applicazioni di transazione ed i file di registro.

Per ottimizzare il processo decisionale, le aziende sfruttano le capacità analitiche del DWH per ottenere importanti insight sul business dai loro dati. Grazie a queste funzionalità, un data warehouse genera un record storico nel tempo che può rivelarsi molto utile per figure come i data scientist e i business analyst e, inoltre, può essere considerato un'unica fonte attendibile di dati aziendali.

Un tipico data warehouse include solitamente i seguenti elementi (18):

- Un database relazionale per archiviare e gestire i dati;
- Una soluzione di estrazione, trasformazione e caricamento (processo ETL) per preparare i dati all'analisi;
- Funzionalità di analisi statistiche, reporting e data mining (estrazione di informazioni implicite precedentemente sconosciute ed esplorazione di grandi quantità di dati al fine di scoprire pattern significativi);
- Strumenti di analisi del cliente per visualizzare ed esibire i dati agli utenti aziendali;

- Altre applicazioni analitiche all'avanguardia che producono informazioni grazie ad algoritmi di data science e artificial intelligence (AI), che permettono ulteriori tipi di analisi dei dati su larga scala;
- Le aziende possono optare anche per soluzioni che uniscono elaborazione delle transazioni, machine learning ed analytics in real time tra DWH e data lake, senza la complessità, la latenza, i costi e i rischi di duplicazione del processo ETL (estrazione, trasformazione e caricamento).

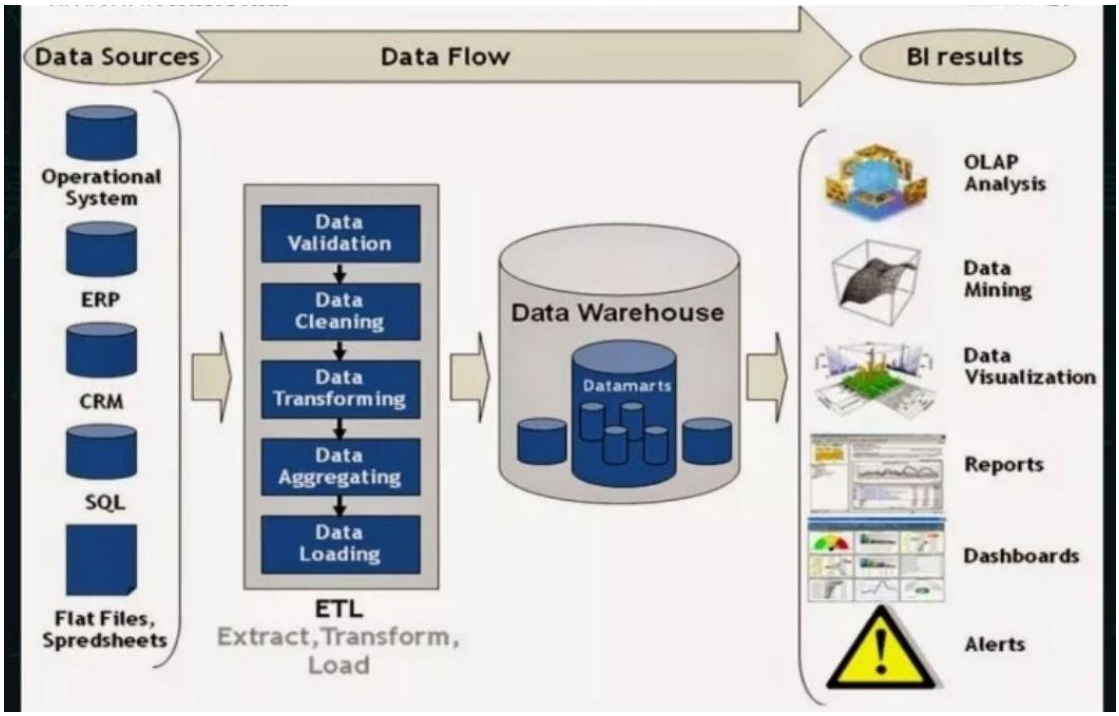


Figura 2.3: Data Warehouse
(28)

2.4.2 Vantaggi di un Data Warehouse

Il beneficio maggiore derivante da un data warehouse è dunque il fatto che quest'ultimo permette alle aziende di analizzare enormi volumi di dati di diversa tipologia, di ricavarne un valore significativo e infine di conservare un record storico.

Essenzialmente, 4 caratteristiche permettono ai data warehouse di procurare questo vantaggio:

- Orientati agli oggetti: possono analizzare dati inerenti ad un particolare argomento o area funzionale;
- Integrati: i data warehouse creano coerenza tra diversi tipi di dati provenienti da fonti differenti;
- Non volatili: una volta che i dati si trovano in un data warehouse, essi sono stabili e non vengono modificati;
- Variante temporale: l'analisi del data warehouse esamina il cambiamento nel tempo.

In modo che soddisfi una varietà di richieste più o meno specifiche, un data warehouse offrirà agli utenti finali una flessibilità sufficiente per ridurre la quantità di dati per un esame più approfondito. Inoltre, nel caso in cui sia ben progettato e strutturato, un DWH eseguirà le query molto più velocemente.

2.4.3 Architettura di un Data Warehouse

In generale, l'architettura di un data warehouse è specifica in base al contesto/azienda in cui viene utilizzata, ma è possibile determinare alcune delle caratteristiche comuni (18):

- Semplice: tutti i data warehouse condividono una struttura di base dove dati di riepilogo, dati non elaborati e metadati sono archiviati nel repository centrale, il quale è alimentato da fonti di dati da un lato ed è accessibile da parte degli utenti finali ai fini di analisi, reporting e data mining dall'altro;
- Semplice con un'area di gestione temporanea: come già descritto in precedenza, i dati devono essere puliti ed elaborati prima di essere messi nel DWH. Per facilitare la preparazione, molti data warehouse aggiungono un'area di gestione temporanea per i dati (nonostante tale operazione possa essere eseguita a livello di programmazione);
- Hub e spoke: quando i dati sono pronti per il loro impiego, vengono trasferiti nei data mart più adatti (database strutturati definiti in base a contesti business specifici). Attraverso il loro inserimento tra il repository centrale

e gli utenti finali, per un'azienda è possibile personalizzare il proprio data warehouse in modo da poter condurre e dirigere differenti linee di business;

- Sandbox: aree private, protette e sicure che permettono alle aziende di esplorare velocemente e informalmente nuovi set/metodi di analisi di dati, senza doversi uniformare alla compliance ovvero alla conformità a determinate norme, regole o standard.

Col passare del tempo ha preso piede l'enterprise data warehouse (EDW), attraverso il quale è stato possibile sviluppare un maggior valore incrementale da poter offrire alle aziende.

Funzionalità	**Valore aziendale**
Reporting transazionale	Procura informazioni relazionali per generare snapshot (catture di stato di un oggetto) delle performance di un'impresa
Analisi approfondita, query ad hoc, strumenti di BI	Espande le funzionalità per insight più dettagliati e analisi più potenti
Previsione delle performance future (data mining)	Sviluppa visualizzazioni e business intelligence predittiva
Analisi tattica (spaziale, statistica)	Propone scenari "what-if" per prendere decisioni pratiche basate su analisi più complete
Archivia molti mesi o anni di dati	Archivia i dati solo per settimane o mesi

Tabella 2.1: Funzionalità e valore aziendale dell'EDW

2.4.4 Quando usare un Data Lake o un Data Warehouse

Un fattore comune che caratterizza sia un data lake sia un data warehouse è il fatto che entrambi vengono adoperati per enormi quantità di dati derivanti da sorgenti differenti. La preferenza nell'impigare uno piuttosto che l'altro dipende dal modo in cui l'azienda prevede di sfruttare i dati. Di seguito vengono mostrate le situazioni migliori in cui utilizzare i due tipi di repository.

Data lake: inizialmente archivia grossi volumi di dati non filtrati da usare in un secondo momento per un'obiettivo specifico. Possibili esempi di dati non elaborati gestiti all'interno di un data lake sono quelli derivanti da sorgenti di app per dispositivi mobili, social media, dispositivi IoT, etc. Solo nel momento dell'effettiva analisi sono ricavate struttura, integrità, selezione e formato dei differenti set di dati. Ad esempio, un data lake potrebbe essere l'opzione corretta nel caso in cui un'azienda abbia bisogno di storage a basso costo per dati non formattati e non strutturati (ricevuti da più sorgenti) che si vogliono adoperare per qualche scopo in futuro.

Data warehouse: sono destinati fondamentalmente all'analisi dei dati e alla conseguente progettazione della reportistica. In tale contesto, i dati sono già pronti quindi sono stati raccolti, contestualizzati e trasformati con lo scopo di produrre insight. Perciò un DWH potrebbe essere l'opzione corretta ogniqualvolta le aziende necessitino di elaborazioni di analytics avanzate. Il processo attraverso il quale si decide quali dati considerare o meno nel data warehouse viene definito "schema on write".

Ingestionare, pulire e trasformare i dati prima che essi vengano archiviati in un DWH può richiedere molto tempo ed è dunque un processo articolato che nega spesso di acquisirli in maniera rapida. Invece, con un data lake si può cominciare nell'immediato a raccogliere i dati e stabilire poi in un secondo momento come usarli.

Tenendo a mente la struttura predefinita di un DWH, quest'ultimo viene impiegato maggiormente dagli utenti aziendali che conoscono fin dall'inizio i dati utili per la reportistica standard.

Invece, le figure professionali che svolgono le ricerche usufruendo dei dati applicando di volta in volta filtri e analisi più avanzati (e di conseguenza coloro che sono più propensi ad adoperare un data lake) sono i data scientist e gli analisti.

Altre differenze riscontrabili tra data lake e data warehouse si individuano a livello di hardware per lo storage e a livello di costi. Generalmente, un DWH può essere oneroso mentre un data lake (dal momento che usa hardware di largo consumo) riesce ad essere più economico nonostante la dimensione elevata.

2.5 Differenze tra approccio batch e approccio real time

2.5.1 Definizione elaborazione batch

L'elaborazione batch è un sistema per mettere in pratica grossi volumi di lavoro sui dati di tipo ripetitivo, progettato per essere un processo gestito con un'interazione minima o nulla da parte dell'utente, perciò quasi del tutto automatizzato (avendo pur sempre a disposizione risorse di elaborazione sufficienti). (2)

L'elaborazione batch è un modo straordinariamente congruo e adatto per elaborare ingenti quantità di dati in poco tempo. Una volta avviato il dispositivo con il quale si sta lavorando, esso si arresta solo nel momento in cui scopre un errore o un'anomalia, avvisando il personale o il dirigente competente.

2.5.2 Quando viene impiegata l'elaborazione batch in ambito aziendale

L'elaborazione batch presenta una serie di vantaggi, ma è ideale nelle aziende in cui (1):

- Vi è un processo che non deve essere affrontato in maniera tempestiva e le informazioni in real time non sono necessarie;
- C'è bisogno di elaborare elevate quantità di dati;
- Vi è un periodo di tempo in cui un computer o un sistema sono inattivi;
- Un processo non richiede l'intervento di personale ed è ripetitivo.

Un esempio consono all'elaborazione batch fa riferimento, ad esempio, al come viene effettuata la fatturazione all'interno delle società di carte di credito. Quando la carta di credito è ricevuta dai clienti, essi ricevono una fattura che non è separata per ogni transazione bensì è relativa all'intero mese. Questo è dovuto al fatto che tutte le informazioni vengono raccolte durante tale periodo ma vengono elaborate in una specifica data tutte in una volta.

Al giorno d'oggi le transazioni vengono generalmente elaborate nell'immediato. Tuttavia, un tempo le banche utilizzavano l'elaborazione batch alla fine di ogni giornata per non occupare risorse di calcolo nei momenti di picco.

Un altro esempio di elaborazione batch (che è tra i più familiari) è il sistema di posta elettronica. Per evitare errori, l'utente ha il tempo di eliminare o aggiornare un'e-mail prima che venga inviata, ad esempio quando si dimentica di includere un allegato.

Questo è dovuto al fatto che la maggior parte dei programmi è in grado di memorizzare i messaggi al suo interno per un preciso periodo di tempo dopo l'invio, e quindi di inviarli sfruttando questo tipo di elaborazione.

2.5.3 Motivi per utilizzare l'elaborazione batch

L'elaborazione batch ha subito diverse modifiche nel corso degli anni. Oggi i dati batch non sono solo un processo di fine giornata o notturno. L'elaborazione non necessita di una connessione a Internet e può essere eseguita in modo asincrono.

In pratica, questi batch possono essere eseguiti in background in qualsiasi momento opportuno senza interrompere i processi vitali. Tuttavia, con l'enorme potenza di calcolo e il cloud computing di oggi, vi sono ancora ottime ragioni per utilizzare l'elaborazione batch.

2.5.4 Vantaggi dell'elaborazione batch

Dopo aver dato definito l'elaborazione batch e i contesti in cui è preferibile il suo impiego, è il momento di passare a descrivere i vantaggi veri e propri che essa può portare con sè (1):

- Velocità e risparmio sui costi
  Non sono richiesti interventi manuali poiché l'elaborazione batch è ampiamente automatizzata. L'automazione limita i costi operativi e incrementa la velocità di elaborazione delle transazioni e dei dati. Se necessario, le organizzazioni possono stabilire un ordine di priorità nell'elaborazione dei dati;

- Precisione
  Eliminando le persone dal processo non vi sono errori umani e si risparmia

così tempo e denaro, con il risultato di dati più precisi e utenti finali più soddisfatti;

- Funzionalità offline
  I sistemi di elaborazione batch funzionano offline. Alla fine della giornata lavorativa questo strumento è ancora in funzione. Per evitare di sovraccaricare il sistema e di interrompere le attività quotidiane, i responsabili possono controllare quando viene avviato un processo;

- Imposta e dimentica
  Non c'è bisogno di accedere, verificare o regolare nulla poichè il sistema è automatico. In caso di problemi viene inviata una notifica al personale competente. Per il resto, si tratta di una soluzione completamente indipendente di cui tutti possono fidarsi;

- Mantenere la semplicità
  Non è richiesto un supporto continuo al sistema, né un inserimento aggiuntivo di dati e neanche un software specializzato. Una volta che il sistema è attivo e funzionante, non c'è bisogno di nessuna manutenzione e si tratta di una soluzione a bassa barriera per l'elaborazione dei dati;

- Dati accurati per il machine learning e l'intelligenza artificiale
  Una delle maggiori sfide dell'AI è la scarsa qualità e affidabilità dei dati. I data scientist dedicano molto tempo alla pulizia dei dati e alla rimozione di errori e incongruenze. Grazie alla sua natura automatizzata, l'elaborazione batch evita totalmente gli errori nei dati. Quando viene riscontrata un'anomalia, questa viene segnalata nell'immediato in modo da poter essere risolta nel più breve tempo possibile garantendo così un risultato finale accurato;

- Migliore utilizzo dei sistemi informatici esistenti
  Poiché l'elaborazione batch può essere automatizzata per essere eseguita quando il sistema raggiunge un certo punto di larghezza di banda, non c'è bisogno di adoperare nuovi sistemi e le risorse esistenti vengono usate in maniera più intelligente. Consentire l'elaborazione dei dati in un momento in cui il sistema è poco sollecitato permette di sfruttare quest'ultimo al massimo.

2.5.5 Problematiche e sfide dell'elaborazione batch

Sebbene l'elaborazione batch sia un'ottima risposta, non è comunque quella giusta per ogni azienda o scenario. Ci sono limitazioni e sfide che potrebbero non renderla la soluzione migliore per ogni organizzazione (1) - (3):

- Formato di dati e codifica
  Nel momento in cui i file usano una codifica o un formato non previsto, si manifestano alcuni dei problemi più complicati da risolvere. Ad esempio, potrebbe essere usata una combinazione di codifica UTF-16 e UTF-8 oppure potrebbero essere inclusi caratteri o delimitatori imprevisti (spazio anziché tabulazione). La logica di caricamento e analisi dei dati deve essere flessibile nell'individuare e gestire queste problematiche;

- Orchestrazione dei periodi di tempo
  In genere è necessaria un'orchestrazione per poter eseguire la migrazione o la copia dei dati nella risorsa di archiviazione dati e nei livelli per la creazione di report. Di solito, i dati di origine sono introdotti in una gerarchia di cartelle che determina gli intervalli di elaborazione, organizzati per anno, mese, giorno, ora e così via. In alcuni casi, i dati possono arrivare in ritardo a seguito di alcuni errori (che dovranno essere ignorati o meno in base alla situazione specifica);

- Formazione
  Tutte le nuove tecnologie richiedono formazione perciò i dirigenti e il personale devono capire come funziona l'intera elaborazione, tra programmazione ed esecuzione delle notifiche di eccezioni ed errori;

- Costi
  Per un'azienda piccola che non dispone di personale addetto all'inserimento dei dati o di hardware sufficiente a sostenere il sistema, i costi di avvio potrebbero non essere gestibili. Esclusivamente per le grosse aziende che elaborano dati continui e voluminosi, l'implementazione dell'elaborazione batch permette di risparmiare tempo e denaro sulla manodopera.

2.5.6 Elaborazione in tempo reale / real time analytics

A differenza dell'elaborazione batch, con l'espressione real time analytics si intende un'analisi dei big data che consente di sfruttare tecnologie e processi attraverso i quali i dati devono essere misurati, gestiti e analizzati in tempo reale non appena entrano nel sistema, permettendo visualizzazioni, comprensioni e approfondimenti rapidi alle aziende.

Tali sistemi sono reattivi e vengono adoperati quando la tempistica è di fondamentale importanza. In questo modo, agire sugli stessi dati consente di perfezionare i flussi di lavoro e il supporto delle vendite e del marketing in modo più efficace.

Inoltre, tali strumenti procurano informazioni sul comportamento ma soprattutto sulle esigenze dei clienti, evidenziando le principali tendenze del mercato e stando così al passo con la concorrenza. (43)

Esempi di elaborazione in tempo reale (87):

- Sistemi di servizio al cliente;
- Sportelli bancomat bancari;
- Flusso di dati;
- Sistemi radar.

2.5.7 I principali vantaggi della real time analytics

In precedenza sono stati discussi i benefici e i vantaggi che l'elaborazione batch può portare con sè. Ovviamente non è da meno la real time analytics, che dalla sua offre miglioramenti su diversi aspetti (43):

- Riconoscimento degli errori
  Se si segue un approccio real time si è sempre in grado di rimanere informati e aggiornati. Contemporaneamente, è possibile vedersi consegnato un avviso ogni volta che il concorrente sta modificando la strategia. La conoscenza di questo metodo assiste le aziende ad essere più reattive su eventuali errori, incrementando di conseguenza la loro efficienza operativa;

- Tempo di reazione aumentato
  Per poter procedere velocemente di fronte ad una situazione urgente ed imminente e sottrarsi perciò a perdite e ricadute, l'elaborazione in tempo reale permette di intercettare i cambiamenti repentini, come ad esempio i cambiamenti nel comportamento d'acquisto dei consumatori o le transizioni di mercato;

- Aumenta il tasso di conversione e i profitti
  Per incrementare il tasso di conversione di un'azienda è necessario che quest'ultima possa migliorare il servizio attraverso il monitoraggio costante dei movimenti dei propri prodotti, il che induce ad una rapida individuazione dei possibili guasti.

  Un servizio che procura informazioni sempre aggiornate sui clienti di un'azienda (dati demografici, numero di visualizzazioni degli annunci sul web, etc), che intercetta i bisogni specifici e che in aggiunta prende decisioni migliori e più rapide, è la customer-care in tempo reale.

  Ciò permette il perfezionamento delle strategie di prezzo e di targeting, metodo strategico volto ad individuare (tramite un processo di segmentazione di mercato) il settore di pubblico ideale per il proprio business;

- Migliore conoscenza delle vendite e monitoraggio del cliente
  Il controllo costante del comportamento d'acquisto di un cliente facilita l'acquisizione di una conoscenza dettagliata sull'andamento delle vendite, consentendo di far comprendere all'azienda quale categoria merceologica sta performando meglio sul mercato;

- Risparmio sui costi
  Vengono limitate potenziali inefficienze e viene ridimensionato in maniera più corretta il capitale umano, traducendo ciò in un maggiore risparmio sui costi;

- Apportare modifiche alla campagna di marketing in tempo reale
  L'approccio real time è vitale sia per la buona riuscita di tutte le iniziative di marketing sui social media sia per la realizzazione di nuovi strumenti volti a migliorare la campagna.

Le aziende che usufruiscono tuttora di dati storici si trovano in netto svantaggio rispetto a quelle imprese che sfruttano questa metodologia/tipologia di dati poichè possono intuire e rendersi conto del comportamento dei clienti e supportare l'efficienza di risparmio dei costi.

2.5.8 Near real time

A differenza dell'approccio real time, l'elaborazione near real time avviene non immediatamente ma comunque in maniera piuttosto rapida (si può parlare di ore o addirittura di minuti). Ciò avviene nel momento in cui la velocità è importante ma non fondamentale, quindi quando il tempo di elaborazione in minuti è accettabile al posto dei secondi.

Un esempio familiare riguarda i processi in background su piattaforme come i siti di social media, che portano a considerare un post ma allo stesso tempo a rimuoverlo dopo pochi istanti, se a livello di norme del sito vengono rilevate alcune violazioni.

Esempi di elaborazione quasi in tempo reale (87):

- Monitoraggio dei sistemi IT;
- Elaborazione delle transazioni finanziarie;
- Elaborazione dei dati del sensore.

2.6 Strumenti di BI

2.6.1 Definizione di Business Intelligence

La business intelligence (BI) è il settore informatico che prevede di raggruppare informazioni valide e proficue per guidare le decisioni di business. Inoltre, questa espressione definisce sia un insieme di tecnologie e operazioni per la raccolta, integrazione e analisi di dati sia tutto ciò che concerne la presentazione delle informazioni estratte dal procedimento stesso.

L'obiettivo è ottimizzare i processi aziendali indirizzando il processo di decision making. Per perfezionare la performance di un'azienda bisogna compiere diverse attività di intelligence:

- Raccolta, analisi e gestione di dati;
- Utilizzo di software per gli interventi di BI;
- Creazione di output con le informazioni strategiche ottenute nel processo.

Operando in tale maniera, le scelte strategiche saranno data driven (guidate dai dati) e di conseguenza verrano prese decisioni migliori per le aziende da parte dei professionisti del settore.

2.6.2 Obiettivo della Business Intelligence

Grazie all'utilizzo strategico della business intelligence si elimina quasi del tutto la soggettività di coloro che sono incaricati a lavorare sui dati, dunque la gestione delle decisioni dell'impresa non può essere condizionata dalle impressioni e sensazioni personali.

Per ottenere un'analisi strutturata e definitiva ed una produzione di qualità di reportistica, i software di BI assistono le aziende nell'integrare dati provenienti da differenti fonti in una vista unica, il che è un fattore imprescindibile.

Tra queste fonti di dati, bisogna considerare ad esempio (80):

- I sistemi di CRM (customer relationship management);
- Tool di marketing analytics;
- Dashboard sulle performance delle vendite;
- Supply chain information systems (strumenti che tengono sotto controllo tutti i flussi in entrata e in uscita di una società);
- Dati e metadati provenienti dagli applicativi di customer care;
- Data wharehouse (in cui le informazioni sono trasformate anche mediante strumenti di data mining).

Per essere efficienti di fronte ad ogni problema dell'azienda, è necessario possedere dashboard unificate e fare reportistica in tempo reale, attività possibili grazie alla BI che confluiscono i dati in un unico luogo (in cui è possibile analizzarli e studiarli).

2.6.3 Importanza della BI per le aziende

Tra i maggiori benefici che le aziende ottengono dalle operazioni di business intelligence ci sono (80):

- Individuazione di opportunità future per migliorare i fatturati;
- Identificazione di criticità, come rallentamenti nei processi aziendali;
- Raccolta dati sui clienti e analisi della customer experience;
- Monitoraggio delle performance;
- Ottimizzazione del budget;
- Controllo dei trend di mercato e dei competitor;
- Integrazione di informazioni raccolte da sorgenti disparate, per mantenere una visione più ampia possibile.

Figura 2.4: Business Intelligence (21)

2.6.4 Strumenti di Business Intelligence

Alcuni dei tool di BI messi a disposizione dell'utente rientrano nella categoria a pagamento mentre altri risultano open source (progetti aperti a cui contribuiscono sviluppatori da tutto il mondo e che possono essere adattati secondo le specifiche esigenze di un'impresa).

In generale, essi consistono in tutte quelle tecnologie e applicativi software programmati per compiere le operazioni di raccolta, integrazione, analisi e output. Eccone alcuni (7):

- Data mining: uso di database, statistiche e apprendimento automatico per individuare le tendenze in vasti set di dati;
- Elaborazione di report: condivisione delle analisi dei dati con i soggetti interessati, affinché possano prendere decisioni e trarre le dovute conclusioni;
- Metriche e benchmarking delle prestazioni: usufruendo di dashboard personalizzate, viene effettuato un confronto tra i dati storici e i dati attuali per monitorare le prestazioni rispetto agli obiettivi;
- Analisi descrittiva: utilizzo di analisi dei dati preliminari per capire cosa è successo;
- Esecuzione delle query: interrogazione dei dati con specifiche domande, per cui la business intelligence estrae le risposte dai set di dati;
- Analisi statistica: partendo dai risultati dell'analisi descrittiva, esplorazione addizionale dei dati che sfrutta le statistiche (per esempio in relazione a come e perché si sia verificata una particolare tendenza);
- Visualizzazione dei dati: trasformazione dell'analisi dei dati in rappresentazioni visive come grafici, diagrammi e istogrammi;
- Analisi visiva: esplorazione dei dati mediante le rappresentazioni visive, per comunicare informazioni velocemente e seguire il flusso dell'analisi;
- Preparazione dei dati: compilazione di differenti fonti dati (individuando misure e dimensioni) e preparazione per l'analisi dei dati.

2.6.5 I vantaggi nell'utilizzo di un software di Business Intelligence

L'impiego di software e tecnologie di Business Intelligence porta con sè una vasta gamma di benefici. Di seguito sono elencati i più rilevanti (6):

- Decisioni basate sui dati
  In passato, per acquisire informazioni c'era bisogno di molto più tempo poichè esclusivamente gli specialisti riuscivano a comprendere e interrogare i dati. Oggigiorno, i software di BI permettono di possedere in real time una visione completa dei dati aziendali significativi. Per avere più tempo da occupare per l'analisi dei dati (ai fini di prendere delle decisioni), bisogna perdere il minor tempo possibile per la fase di reportistica;

- Cruscotti intuitivi
  Grazie alla business intelligence molte organizzazioni sono riuscite ad incrementare l'efficacia dei propri report limitando di gran lunga i tempi di analisi. Non è necessario che l'utente abbia una particolare formazione poichè questi software permettono di estrapolare i dati da più sorgenti e possedere report veloci di facile consultazione;

- Maggiore efficienza organizzativa
  È possibile individuare le aree che concedono maggiori possibilità grazie alla visione a 360° di cui possono disporre i manager e compararle con i risultati dell'organizzazione più vasta;

- Migliorare l'esperienza dei clienti
  Le aziende sono in grado di individuare in maniera rapida cosa manca ai propri servizi, apportare le giuste modifiche per perfezionare i processi di assistenza e incrementare la soddisfazione dei clienti sfruttando i loro feedback e le conoscenze sul loro comportamento;

- Maggiore soddisfazione dei dipendenti
  I diversi reparti possono aver accesso in modo semplice ai propri dati senza dover contattare figure specifiche come analisti o IT. Anche utenti senza una particolare formazione sono in grado di ispezionare i dati utili al loro lavoro.

Ciò è possibile grazie alla scalabilità offerta da questi tool che permettono inoltre la condivisione delle dashboard con i dipendenti, il che facilita il loro coinvolgimento e la loro soddisfazione;

- Migliore qualità dei dati
  I software di BI assistono la pulizia dei dati (che possono presentare imprecisioni e discrepanze) per far avere all'interno dell'azienda un quadro ampio e completo di ciò che sta accadendo;

- Aumento del vantaggio competitivo
  Per comprendere in modo chiaro il mercato e le prestazioni dell'azienda, è necessario che i tool di BI permettano di analizzare i dati provenienti da sorgenti disparate (sia interne che esterne). Con tali strumenti si possono incrementare i guadagni prendendo le decisioni strategiche ottimali che anticipano le necessità degli utenti.

2.6.6 La differenza tra BI tradizionale e BI moderna

In passato, c'era un problema di fondo per la BI tradizionale: si adoperavano report statici per dare una risposta alla maggior parte delle domande di analisi ma nel momento in cui chiunque avesse avuto un ulteriore questito sul report ricevuto, la sua richiesta veniva messa in fondo alla coda di reporting e il processo doveva ripartire dall'inizio.

Si trattava di un approccio top-down in cui la BI era gestita dall'organizzazione IT. Pertanto, gli utenti non erano in grado di trarre profitto dai dati attuali per prendere decisioni poichè i cicli dell'attività di report erano lenti, frustranti e deludenti. Tuttavia, per rispondere a query statiche e per le elaborazioni standard di report è ancora comune usare la BI tradizionale.

Tutt'altro discorso per la BI moderna: con il software opportuno e anche con poco preavviso, diversi livelli di utenti possono visualizzare i dati e dare una risposta alle proprie domande personalizzando le dashboard e creando report. Nonostante i reparti IT siano ancora basilari per la gestione dell'accesso ai dati, la versione moderna della business intelligence è decisamente più accessibile e interattiva.

2.6.7 Principali software di Business Intelligence

Dopo aver descritto nel dettaglio gli obiettivi, gli strumenti e i vantaggi della Business Intelligenge, è il momento di stilare la lista dei software maggiormente impiegati sul mercato (6) - (77):

1. Microsoft Power BI
   Power BI è un programma di BI che permette la visualizzazione dei dati in modo interattivo. La sua interfaccia (Power BI Desktop) permette una semplice consultazione di tutti i dati.

   È disponibile (seppur con funzionalità ridotte) in una versione gratuita oppure in una versione Pro che si rivolge alle aziende. Dotato di una vasta documentazione e aggiornato in modo continuo, esso si compone di un servizio basato sul cloud. I punti a favore di Microsoft Power BI si possono riassumere sicuramente nella sua ampia diffusione, nel suo costo limitato e nella sua facilità d'uso.

   Il servizio cloud Power BI è decisamente ricco nelle sue capacità (tra cui un set ampliato di analitiche aumentate e capacità di machine learning automatizzate) e può disporre di svariate funzionalità di prodotto. Inoltre, conta su una ragguardevole visione e capacità d'esecuzione dello sviluppo, con numerosi nuovi rilasci e funzioni innovative;

2. Tableau
   Integra un'elevata quantità di sorgenti di dati e garantisce differenti possibilità di visualizzazione (attraverso dashboard), esplorazione e analisi dei dati in maniera del tutto intuitiva e interattiva. Si tratta di uno strumento facile e potente da adoperare, tra l'altro uno dei primi software di BI utilizzati.

   Inoltre, i processi d'analisi hanno subito un'accelerazione dovuta all'automazione dei compiti di routine e alla facilitazione della scrittura delle formule, grazie agli ultimi rilasci delle soluzioni di BI e Analytics;

3. Qlik Sense
   Dispone di tutte le funzionalità di un software di BI, dall'analisi avanzata per ogni necessità aziendale fino alla visualizzazione in cruscotti intuitivi e reportistica.

Malgrado abbia uno slancio di mercato limitato rispetto agli altri due software, è comunque molto apprezzato per la sua interfaccia semplice e intuitiva ed, inoltre, fa rientrare tra i suoi punti forti l'elevata flessibilità di implementazione e la vision di prodotto in termini di tecniche di intelligenza artificiale e machine learning.

Inoltre, con l'entrata in scena di nuovi competitor in ambito cloud sono da considerare anche:

4. QuickSight
   È un servizio di Business Intelligence nativo per il cloud e serverless (gli sviluppatori possono creare e gestire applicazioni senza doversi curare dell'architettura sottostante).

   Connettendosi con QuickSight, è possibile dimensionare l'intero set di dati, creare pannelli di controllo personalizzabili, sfruttare le integrazioni del Machine Learning per informazioni precise, abilitare una Business Intelligence self-service per tutti, integrare dei servizi AWS nativi e pagare in base all'uso. Inoltre, sono garantite sicurezza, governance e conformità integrate; (8)

5. Looker
   Piattaforma di BI incorporata in Google Cloud (non limitata a sfruttare un'unica interfaccia o un unico ambiente cloud) con la quale si è in grado di andare al di là delle semplici dashboard o report. Ottimizza il processo decisionale analizzando, visualizzando e intervenendo sugli insight aggiornati su cui gli utenti possono e devono fare affidamento.

   La moderna architettura in-database che assegna metriche con dati in real time, consente di attuare decisioni più consce e consapevoli. Inoltre, per procurare dati regolamentati all'interno dell'azienda, flussi di lavoro basati sui dati e applicazioni personalizzate, Looker fa uso di una piattaforma incentrata sulle Application Programming Interface o API (l'insieme di definizioni e protocolli necessari per la creazione e l'integrazione di applicazioni software).

   Con tale piattaforma è possibile migliorare i processi aziendali e incrementare l'efficacia, automatizzando attività che prima erano ripetitive e monotone. Con

un modello dati centralizzato non c'è bisogno di uno storage in-memory, infatti è LookML che si occupa autonomamente di applicare le regole di business direttamente alla base dati sottostante. (9)

Capitolo 3

Tecnologie Cloud

3.1 Confronto tra Data Platform e Modern Data Platform

3.1.1 Modern Data Platform: necessità di una piattaforma moderna

Con la diffusione dell'IoT (Internet of Things), dei Big Data, del cloud e più recentemente dell'intelligenza artificiale, il mondo dei dati si è trasformato in maniera radicale.

Nello specifico, sono incrementate le fonti dati con nuovi formati e interfacce, ma allo stesso tempo sono state modificate anche le destinazioni con l'inserimento di servizi di Analytics tra cui il data lake. Inoltre, la necessità di avere sempre a disposizione il dato (con il bisogno di effettuare azioni e trasformazioni sempre più in tempo reale o quasi) non è più la stessa.

Per queste ed altre ragioni, ad oggi si parla di Modern Data Platform ovvero una piattaforma moderna di analisi dati. A causa dell'ingente aumento delle fonti dati e quindi dell'elevata ricchezza informativa, diverse aziende si trovano a dover gestire in maniera completamente differente l'approccio che stava alla base dei processi e delle attività del passato, quindi la Modern Data Platform ha l'obiettivo di aggiornare e rivoluzionare il sistema di analisi odierno.

3.1.2 Definizione di una Modern Data Platform

Un primo fattore da tenere in considerazione riguarda la tipologia dei dati: la modern data platform si basa sia sull'archiviazione di grossi volumi di dati che provengono da svariate sorgenti in formati differenti, sia sull'ottenimento di informazioni vantaggiose che permettono di rendere le decisioni aziendali ottimali (i primi data warehouse erano invece incentrati più sull'elaborazione dei dati).

Un secondo fattore da valutare consiste nella gestione dei volumi di dati: quest'ultima è piuttosto complicata da condurre con i data warehouse tradizionali. Ciò è dovuto al fatto che oggigiorno esistono ingenti quantità di dati strutturati, semi strutturati e non strutturati.

Per stare al passo con i tempi, la diffusione dei data lake (utili per archiviare dati di qualsiasi tipologia che provengono da più sorgenti) è venuta in soccorso alle piattaforme tradizionali. Tuttavia, questi due ambienti di analisi rimanevano sempre ben divisi. Il modo per unificarli è stato ottenuto proprio con l'introduzione della modern data platform.

È esattamente questo lo scopo primario della piattaforma: con essa si possono gestire in modo opportuno i dati multi-strutturati convogliando tutte le esigenze di analisi in un unico ambiente. A differenza di un data lake, la modern data platform valorizza i concetti di certificazione del dato e di governance e, oltre ad essere una piattaforma di analisi completa, può essere anche vista come un data warehouse organizzato dei dati.

3.1.3 Utilizzo di una Modern Data Platform

La modern data platform può gestire sorgenti dati relazionali e non relazionali sfruttando tecniche di streaming in tempo reale e uniformando i dati on prem con quelli su cloud. La modern data platform migliora e facilita il lavoro in diversi ambiti.

Inoltre, essa procura un motore analitico per l'analisi predittiva e l'esplorazione interattiva di dati aggregati, quindi rende il lavoro di un'azienda più semplice e ottimizzato. Un ruolo importante è ricoperto dal linguaggio SQL (servizio di query) che permette alla piattaforma di interrogare in modo semplice i dati relazionali e non relazionali.

Essa sfrutta il processo ETL (estrazione, trasformazione e caricamento) per l'arricchimento dei dati che vengono in seguito preparati attraverso tecniche di Big Data: ingestion oppure processo ELT (estrazione, caricamento e trasformazione).

La modern data platform deve sostenere una nuova tipologia di analisi, con modelli analitici predittivi che diano appoggio e assistenza in real time nel processo decisionale, attraverso figure specializzate che hanno però bisogno di generare analisi in un ambiente condiviso e su dispositivi diversi.

Fondamentalmente, la modern data platform deve sostenere a livello di funzionalità tutti gli strumenti che le aziende possono adoperare per acquisire informazioni dai dati. Ciò racchiude e comprende strumenti self-service che facilitano (attraverso strumenti di BI) l'analisi da parte degli analisti aziendali.

Per riassumere quanto detto finora, ecco un'immagine esplicativa che rappresenta una potenziale modern data platform:

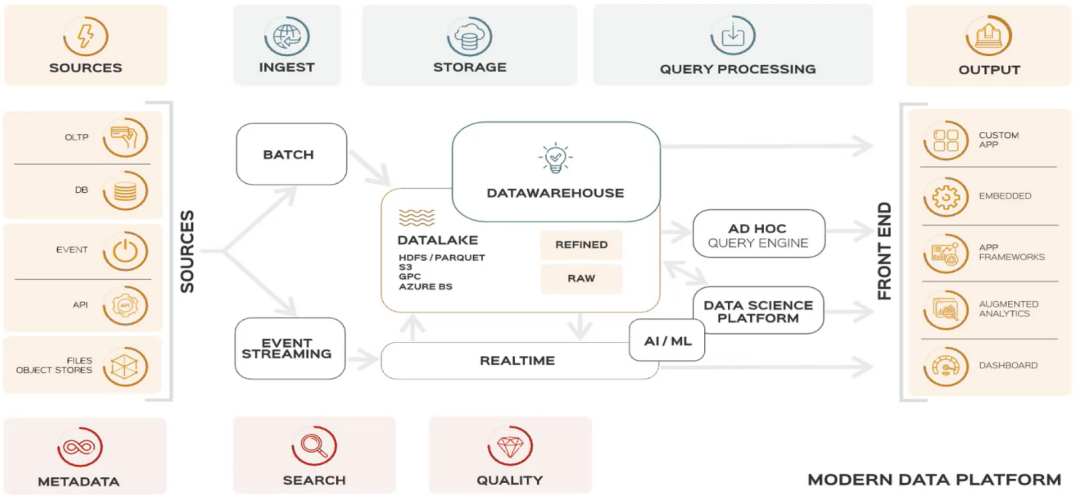


Figura 3.1: Modern Data Platform
(68)

3.1.4 I vantaggi di una Modern Data Platform

L'utilizzo di una soluzione gestita di modern data platform porta con sè numerevoli vantaggi. Tra i principali si evidenziano (40) - (57):

- Ottimizzazione dei tempi
  Netti miglioramenti nei tempi di gestione, nello specifico minor tempo da riservare alla preparazione dati e quindi da poter sfruttare per l'inserimento di funzionalità di analisi avanzate;

- Copertura di tutti gli ambienti di analisi
  La scalabilità fornisce una diversa ampiezza di dati a cui un utente può accedere in base al suo livello;

- Integrazione scalabile dei dati
  Possono essere gestite passo dopo passo tutte le fasi principali tra cui l'ingestion, la convalida, la pulizia e la preparazione dei dati;

- Governance end-to-end
  È possibile individuare e comprendere la qualità delle informazioni classificando i dati secondo etichette automatiche e preimpostate;

- Analisi self-service
  La disponibilità di report utili al business e al processo decisionale è dovuta a dati che si dimostrano ben strutturati e puliti da duplicati o altri errori.

3.1.5 Conclusioni: ragioni per adottare una Modern Data Platform

Affinchè gli utenti aziendali possano avere tutti i dati disponibili e le informazioni necessarie per ottenere gli insights di analisi, essi devono far uso di questa piattaforma di analisi completa. Tale ambiente deve poter incorporare nuove funzionalità in qualsiasi istante quando se ne ha bisogno, senza dover rimettere in discussione l'architettura utilizzata.

Inoltre, questa piattaforma deve stare sempre al passo con le necessità in continuo aggiornamento del mercato, rendendo il livello delle performance richieste duraturo nel tempo. Un ruolo importante spetta al cloud che, oltre a garantire maggior flessibilità nella gestione dei costi e maggior scalabilità delle risorse, garantisce anche un alto livello di servizio.

3.2 Peculiarità e vantaggi dell'uso di tecnologie cloud

Il cloud è la distribuzione attraverso la rete internet di servizi come server, database e software. Un fornitore offre perciò ai suoi utenti servizi di archiviazione, elaborazione e trasmissione dei dati in modalità on-demand.

Al giorno d'oggi si stanno diffondendo a dismisura diverse tecnologie cloud, le quali sono sempre più importanti nei processi di business di qualsiasi settore industriale (privati, start-up, grosse aziende).

L'esigenza di sfruttare tecnologie cloud si è manifestata nel momento in cui ci si è resi conto che in passato c'era bisogno di possedere dei software caricati sul proprio dispositivo o in un server fisico della propria azienda per poter usufruire di programmi o applicazioni. Grazie ad internet, ora è possibile accedere agli stessi programmi da remoto.

3.2.1 Modelli di servizio di cloud computing

Il cloud computing può essere suddiviso in tre principali categorie o modelli di servizio (55):

- Infrastructure as a Service (IaaS): attraverso la rete Internet, procura alle aziende un'infrastruttura IT su base on-demand che comprende server, storage e capacità di rete. Mentre il provider di servizi mantiene l'hardware sottostante, gli utenti sono in grado di gestire il sistema operativo e le applicazioni;

- Platform as a Service (PaaS): ambiente di sviluppo che sostiene il ciclo di vita completo dell'applicazione web. Tale modello permette alle aziende di dedicarsi alla realizzazione e gestione delle loro applicazioni, delegando la manutenzione dell'infrastruttura al provider del servizio cloud. Oltre all'infrastruttura di base, vengono introdotti anche middleware, strumenti di sviluppo, sistemi di gestione dei database e così via;

- Software as a Service (SaaS): in questo modello, il provider controlla l'hardware e il software, togliendo all'utente la preoccupazione per l'aggiornamento e la manutenzione del sistema. Inoltre, agli utenti è permesso connettersi e

usufruire di applicazioni partendo da servizi di posta elettronica fino ad arrivare ad applicazioni innovative di business.

Con l'immagine seguente è possibile capire in modo chiaro le differenze tra i 3 modelli a livello di gestione dei servizi offerti (quindi cosa deve essere gestito dall'azienda e cosa invece può essere gestito dal vendor):

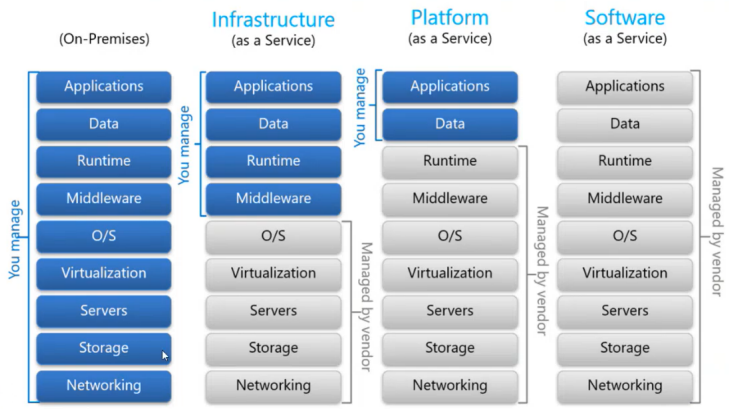


Figura 3.2: IaaS, PaaS, SaaS
(68)

3.2.2 Vantaggi del Cloud

I vantaggi della tecnologia cloud impattano su diversi fronti e sono molto tangibili, variegati e significativi (50):

1. Riduzione dei costi (e aumento dell'efficienza)
   Non è più necessario adoperarsi di server e data center locali, poichè ciò implica personale specializzato nella manutenzione e nella gestione (che a sua volta occupa spazio fisico e consuma elettricità). Nel momento in cui vengono sfruttate tecnologie cloud da parte delle aziende, quest'ultime riducono di gran lunga i costi relativi all'acquisizione di hardware e software.

   In aggiunta, i periodi in cui un sistema informatico non è operativo (per via di guasti, manutenzione o altre cause che scaturiscono i cosiddetti downtime) svaniscono quasi definitivamente e con loro vengono meno i relativi problemi

di inefficienza. Registrare downtime in ambiente cloud è sempre più difficile con il progresso della tecnologia.

Per di più, si pagano esclusivamente le risorse effettivamente utilizzate (pay-as-you-go) a prescindere dal modello di servizio cloud selezionato. In tale maniera, si consente ai team IT di riservare del tempo a un lavoro più strategico, evitando di sovraccaricare ed eseguire l'overprovisioning del data center (tecnica di creazione del file system all'interno dell'SSD che ha come scopo quello di fornire delle celle di riserva);

2. Scalabilità
   Spesso accade che le piccole start-up si ritrovino nel giro di poco tempo a espandersi sull'onda di un rapido successo. Quest'ultimo può essere presto soppresso se le infrastrutture adoperate non sono scalabili e non consentono di incrementare l'operatività in maniera piuttosto spedita.

   In tal senso viene in soccorso il cloud, che tramite una gestione in out-source di diversi settori e tramite l'elasticità dei suoi sistemi, controlla servizi e capacità di calcolo on-demand espandendo velocemente la propria infrastruttura (che allo stesso tempo, dove necessario, potrebbe ritornare con la massima flessibilità ad essere più contenuta, senza drammi relativi agli investimenti passati);

3. Prestazioni
   In termini di prestazioni ma anche di velocità, efficacia e sicurezza, un singolo data center aziendale non può tener testa alla rete di data center forniti da aziende specializzate. Ospitando piattaforme, software e database in remoto, i servizi cloud consentono di svincolare spazio di calcolo sui singoli dispositivi della propria azienda. Perciò, senza investimenti costosi e con grande flessibilità e semplicità, ingenti volumi di risorse di calcolo sono disponibili in tempi brevi;

4. Produttività, in ottica omnichannel
   Tutta una serie di azioni quotidiane nell'ambito IT, tra cui l'assemblaggio e l'organizzazione dei data center locali (con le diverse configurazioni hardware e software), portano via meno tempo con l'introduzione del cloud. Affidarsi all'esterno (ma senza incrementi di costi) a fornitori di servizi Cloud consente

tra l'altro di trarre profitto da know-how (competenze) consolidate, utili per avere delle infrastrutture più conformi al proprio business.

Inoltre, questo tipo di servizi (e di relative interfacce) vanno organizzati nella maniera più efficiente possibile in un'ottica che sia omnichannel, ovvero viene consentito un accesso semplice ai dati e alle piattaforme aziendali con tutti i tipi di apparecchi, dal tipico desktop fino ai dispositivi mobili (che si stanno rivelando sempre più essenziali e imprescindibili come strumenti di lavoro);

5. Affidabilità e sicurezza
   Con il cloud i problemi di affidabilità e sicurezza si esauriscono, in quanto anche in caso di guasti dei dispositivi aziendali tutti i dati sono sempre a disposizione. In ogni caso, sono state attuate procedure di recupero dati ideate per ogni situazione d'emergenza, da parte di tutti quei fornitori di servizi cloud più seri e affidabili.

   Si tratta insomma di rischi che non si possono e devono correre. Infatti, i provider di servizi cloud sono equipaggiati di standard di sicurezza in continuo e costante aggiornamento, che possono assicurare un livello di attenzione inarrivabile ai sistemi di archiviazione e condivisione interni alla singola organizzazione.

   Per esempio, nel contesto cyber security basta un semplice attacco hacker o una grave perdita di dati per compromettere l'immagine di un marchio o di un'azienda e di conseguenza la loyalty (fedeltà e soddisfazione) dei clienti.

3.2.3 Limitazioni del cloud computing

Come qualsiasi tecnologia, oltre ai vantaggi c'è anche ovviamente da considerare il rovescio della medaglia ovvero tutti i vincoli a cui è sottoposto in questo caso il cloud computing (12) - (72):

- Uno svantaggio piuttosto ordinario del cloud fa riferimento alla connessione a internet, la quale se risultasse errata potrebbe non consentire l'acquisizione di informazioni o applicazioni di cui uno necessita.

Invece, dal punto di vista dell'informatica tradizionale, per aver accesso ai dati su server o dispositivi di archiviazione viene sfruttata una connessione cablata;

- Nel caso in cui la propria applicazione esiga un'enorme quantità di dati da trasferire allora il cloud non sarebbe la scelta ottimale da prendere in considerazione.

  Ciò è dovuto al fatto che i benefici derivanti da una elevata capacità di calcolo verrebbero in gran parte messi da parte. Questo problema si riferisce alla latenza;

- Verrebbe evidenziata una complessità d'integrazione con i sistemi esistenti;

- Si manifesterebbe un minor controllo sull'infrastruttura cloud sottostante.

Nonostante il fatto di avere un numero inferiore di svantaggi rispetto alle motivazioni che invece spingono ad adottare un'infrastruttura cloud, bisogna comunque non far passare tutto ciò in secondo piano. È anche vero che valutando in maniera scrupolosa i provider di servizi cloud e i relativi modelli di servizio, la maggior parte di queste problematiche possono essere risolte.

L'opzione di una piattaforma cloud aperta può garantire maggiore libertà e flessibilità nel generare e gestire le risorse di cui uno necessita, integrandole in modo accurato con i servizi che si desiderano.

Bisogna però ricordare che diverse attività rimangono responsabilità del cliente (ad esempio quelle relative alla sicurezza ma anche alla comprensione dei modelli e dei servizi), perciò tanti dei problemi che si verificano durante la migrazione al cloud sono dovuti proprio a questo.

3.2.4 Motivi per passare al cloud computing

Attualmente, quasi tutte le aziende non valutano se eseguire la migrazione al cloud ma valutano cosa dovrebbero migrare. Infatti, tenendo a mente quando detto finora, è innegabile il fatto che i vantaggi nell'impiego del cloud siano maggiori rispetto ai punti a sfavore.

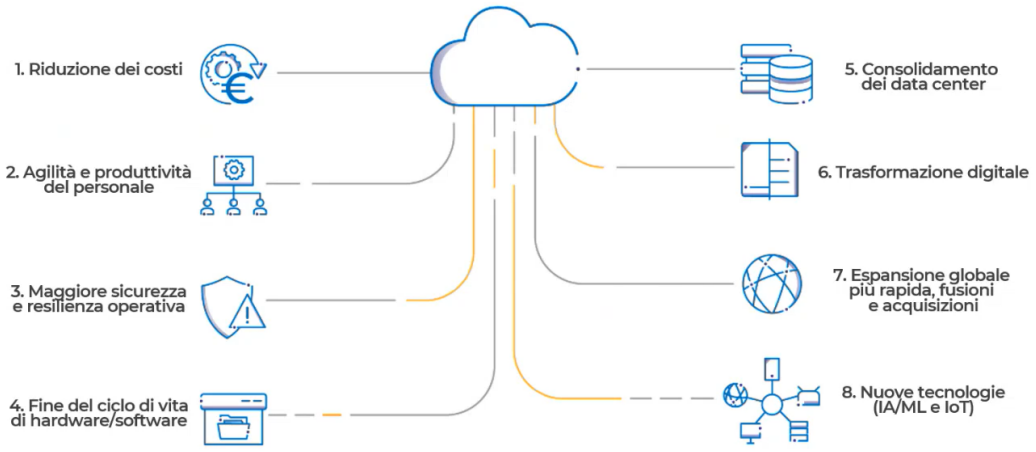


Figura 3.3: Vantaggi del cloud (68)

I vantaggi primari (maggior affidabilità e flessibilità, prestazioni ed efficenza ottimizzate, costi ridotti, innovazione sempre più all'avanguardia grazie all'impiego nelle strategie d'azienda di intelligenza artificiale e machine learning) possono anche tradursi in altri vantaggi associati che possono consentire un incremento di produttività, sostenendo la forza lavoro da remoto.

A seconda del contesto, un'azienda può adottare un approccio ibrido che non vincoli il cloud ad essere uno scenario del tipo o tutto o niente. Infatti, pur agendo nell'ambiente che funziona meglio per l'organizzazione, operando secondo questo schema si manifesta una certa facilità nello sviluppo della capacità e funzionalità dell'infrastruttura esistente.

3.3 Implementazione sistemi data analytics nel cloud

3.3.1 Definizione di Cloud Analytics

Il termine Cloud Analytics si riferisce a un servizio/modello di erogazione per l'hosting che provvede all'analisi e al calcolo dei dati aziendali sfruttando tecnologie cloud. Queste tecnologie gestiscono la memorizzazione e l'elaborazione dei dati.

La Cloud Analytics sta diventando sempre più gettonata e apprezzata dalla maggior parte delle aziende moderne di analisi dati. La migrazione al cloud è per molte organizzazioni l'obiettivo primario da conseguire. Ciò è dovuto in parte alle promesse di incremento della produttività e consistente calo dei costi.

Come detto nel paragrafo precedente, diverse aziende seguono un approccio ibrido di cloud analytics (che può comprendere qualsiasi tipo di Data Analytics o di BI), dunque viene previsto l'impiego di funzioni da eseguire in parte su server on-prem in parte su ambienti cloud.

Invece, una parte delle aziende si sta trasferendo totalmente nel cloud senza adottare nessun approccio ibrido. Ciò elimina il peso della tradizionale esecuzione di analisi (svolta unicamente on-prem) che può portare ad una gestione tutt'altro che economica da parte dell'intera organizzazione. In questo modo, man mano che il business si espande, è possibile scalare il proprio programma di analisi.

3.3.2 3 approcci al Cloud

Dopo aver definito il concetto di Cloud Analytics, è il momento di proseguire andando ad evidenziare i diversi approcci al Cloud. Ce ne sono sostanzialmente 3:

1. Public Cloud: come è intuibile dal termine, in un cloud pubblico i servizi sono disponibili pubblicamente mediante una terza parte come applicazioni, macchine virtuali, capacità di archiviazione e così via. Benchè a volte gli utenti debbano pagare per l'uso o il consumo, di solito i servizi sono offerti in modo gratuito. In questa circostanza, è possibile avere costi ridotti e meno manutenzione poichè i sistemi IT sono condivisi;

2. Private Cloud: a differenza del public cloud, un cloud privato è riservato a utenti prestabiliti di un'azienda/organizzazione. Esso si trova in un data center di proprietà dell'impresa o di un servizio di hosting. Se da un lato esso garantisce una maggiore privacy e sicurezza dei dati, dall'altro lato spesso può essere molto più oneroso. In analogia al public cloud, offre vantaggi tipici come scalabilità e accesso democratizzato;

3. Hybrid Cloud: combinazione dei due approcci precedenti. Per i dati non sensibili l'azienda seleziona una struttura di cloud pubblico mentre per i dati destinati unicamente all'organizzazione stessa si opta per un cloud privato.

3.3.3 Considerazioni per l'implementazione della Cloud Analytics

La Cloud Analytics dispone di una varietà di aspetti e caratteristiche di cui è buona norma esserne a conoscenza. Ci sono diverse considerazioni da fare per la sua implementazione, alcune delle quali sono riportate di seguito (13):

- Potenza di calcolo: utile per assicurare l'acquisizione, la strutturazione e l'analisi dei dati su scala;

- Fonti di dati: una soluzione solida e robusta deve poter acquisire i dati da differenti sorgenti, tra le quali i siti web aziendali, piattaforme di social media, Internet of Things (IoT), app mobili, CRM e così via;

- Modelli di dati: i modelli di dati basati su cloud standardizzano la maniera con la quale i dati si relazionano tra loro e determinano la loro struttura. Inoltre, l'uso di modelli di dati può dare una mano al business trasferendo i dati dall'on-prem al cloud;

- Applicazioni di elaborazione: per elaborare enormi set di dati che provengono da svariate fonti, le aziende necessitano di un framework di elaborazione per i loro ambienti basati su cloud. Queste applicazioni sono sviluppate da diverse realtà presenti sul mercato (ad esempio Google BigQuery, Apache Spark e Hadoop);

- Modelli analitici: c'è bisogno di sviluppare modelli per la Predictive Analytics e altre funzioni analitiche avanzate da eseguire nel cloud;

- Condivisione dei dati: mediante una moderna architettura cloud si è in grado di sostenere un'archiviazione e condivisione dei dati piuttosto semplice.

3.4 Intelligenza Artificiale e Machine Learning

Non esiste una definizione univoca che esprima in modo esaustivo il concetto di intelligenza artificiale, ma si può affermare che essa indica tutti quei sistemi ideati e realizzati dall'uomo in forma di software (ma anche hardware) che stabiliscono le azioni ottimali mediante l'acquisizione e la conoscenza dei dati (strutturati o meno), i quali dovranno essere interpretati per poter estrarre le relative informazioni.

I sistemi di intelligenza artificiale possono imparare e assimilare un modello numerico, utilizzare regole logiche ma soprattutto adattare il loro comportamento analizzando gli effetti che le loro azioni precedenti hanno prodotto sull'ambiente. L'intelligenza artificiale esprime quindi un insieme estremamente vario di sistemi tecnologici in grado di agire in autonomia per soddisfare un particolare obiettivo.

3.4.1 Test di Turing

Per stabilire se una macchina è in grado di pensare come un essere umano oppure no, spesso si chiama in causa il test di Turing (test di AI che è stato ideato nel 1950 dall'omonimo Alan Turing, uno dei padri dell'informatica e uno dei più grandi matematici del XX secolo).

Il test si pone l'obiettivo di capire se una macchina può essere definita intelligente o meno. Per quanto riguarda il suo funzionamento, sono previste due differenti fasi.

Nella prima fase sono coinvolte 3 persone che sono separate in 3 stanze differenti: un intervistatore, un uomo e una donna. L'intervistatore deve indovinare il sesso delle due persone e per far ciò formula loro delle domande scritte.

Le risposte dell'uomo e della donna devono essere trasmesse in modo impersonale o anonimo, al fine di privare l'intervistatore di indizi come l'analisi della grafia o della voce. I partecipanti al gioco ricevono la domanda e rispondono all'intervistatore in forma scritta e in incognito.

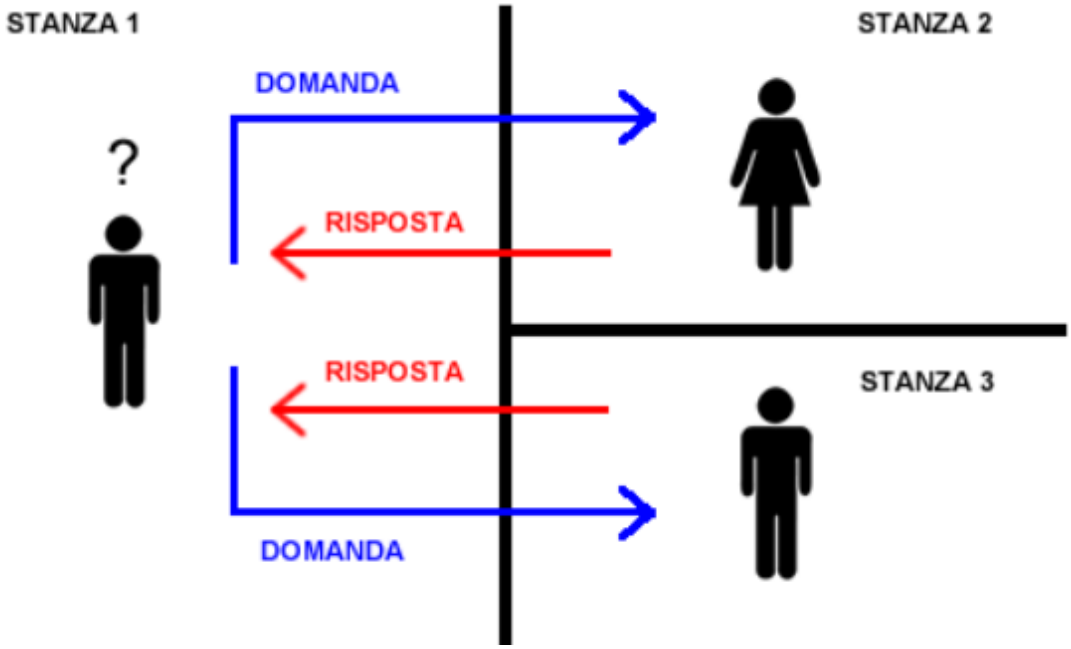


Figura 3.4: Test di Turing prima fase
(76)

La complicazione del problema sorge nel momento in cui ai partecipanti è concesso anche mentire. Infatti, l'uomo e la donna hanno scopi differenti.

Uno dei due è sincero perciò ha lo scopo di agevolare la sua identificazione da parte dell'intervistatore mentre l'altro mente perciò ha lo scopo di far cadere in errore l'intervistatore. Quest'ultimo non è a conoscenza di colui che è sincero e di colui che mente, deve essere in grado di capirlo da sé.

Al termine del gioco l'intervistatore deve stabilire chi è l'uomo e chi è la donna. Ripetendo il gioco N volte (con N sufficientemente grande), l'intervistatore sbaglierà X volte. Il suo tasso di errore sarà pari a $\frac{X}{N} = \alpha$.

Nella seconda fase del test, il processo viene ripetuto nello stesso identico modo ma questa volta si sostituisce la persona che mente con una macchina. In questo caso, l'obiettivo dell'intervistatore consiste nel capire se a rispondere è la persona o il computer.

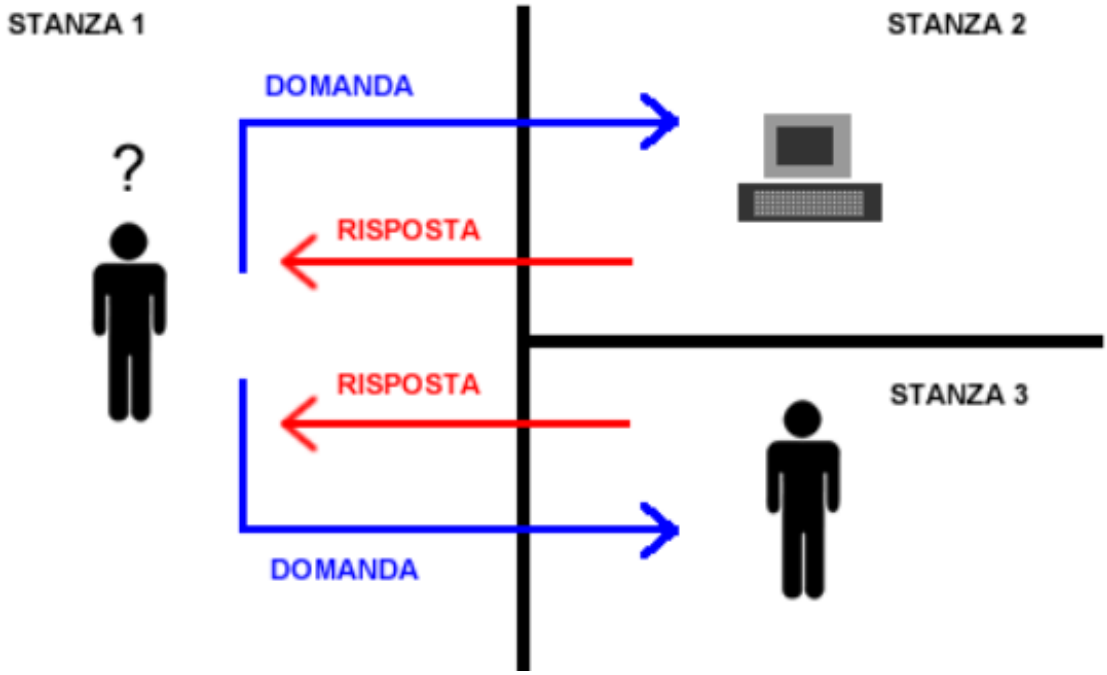


Figura 3.5: Test di Turing seconda fase
(76)

Il gioco si ripeterà N volte (con N sufficientemente grande) e l'intervistatore sbaglierà Y volte, ottenendo un tasso di errore pari a $\frac{Y}{N} = \beta$.

Se α è simile a β, allora il Test di Turing viene superato in quanto la macchina (in tal caso specifico) sarebbe indistinguibile da un essere umano. Col passare degli anni, sono state proposte diverse varianti ma quella appena descritta è la versione originale del test.

3.4.2 Classificazioni e obiettivi dell'intelligenza artificiale

Col passare del tempo si sono diffusi due criteri e punti di vista applicativi ben differenti dell'intelligenza artificiale (nonostante si senta parlare di questo topic quasi unicamente attraverso il suo termine generico). È il caso dell'intelligenza artificiale debole e dell'intelligenza artificiale forte.

Prima di addentrarsi nell'individuazione delle differenze tra AI debole e AI forte, è necessaria una premessa relativa all'obiettivo dell'intelligenza artificiale ovvero riprodurre alcune delle funzioni dell'intelligenza umana. Nello specifico (86):

- Agire umanamente, in modo analogo a quanto farebbe un essere umano nella stessa situazione;
- Pensare umanamente, risolvendo un problema con funzioni cognitive;
- Pensare razionalmente, in modo logico, come l'essere umano fa nei suoi ragionamenti;
- Agire razionalmente, per provare a raggiungere il risultato ottimale sulla base delle informazioni possedute.

Imitare l'uomo non può essere un'operazione univoca ma può essere indirizzata su approcci differenti. Proprio per queste motivazioni, si sono diffusi due filoni ben distinti: da un lato l'intelligenza artificiale debole è ormai ampiamente diffusa mentre dall'altro l'intelligenza artificiale forte si pone un obiettivo ben più complicato da ottenere poichè decisamente più ambizioso.

Figura 3.6: Intelligenza Artificiale
(30)

3.4.3 Intelligenza artificiale debole

Con intelligenza artificiale debole si vogliono indicare tutti quei sistemi ideati e sviluppati dall'essere umano necessari per la risoluzione di determinati problemi,

che possono essere complessi o meno. Il punto chiave è il problem solving, in quanto vengono simulate alcune capacità dell'uomo senza però l'esigenza di voler replicare e comprendere l'intero funzionamento del cervello umano.

Le applicazioni dell'AI debole sono ideali per consigliare all'uomo quali decisioni seguire ed attuare, e sono incentrate sull'apprendimento automatico per generare sistemi capaci di simulare scenari e sostenere l'uomo (garantendogli più insight possibili per rafforzare la sua scelta nello svolgimento di determinati incarichi).

Lo scopo è sempre quello di fornire una risposta ad una necessità pratica nel miglior modo possibile, che non corrisponde sempre con quello che l'uomo avrebbe fatto nella stessa condizione.

3.4.4 Intelligenza artificiale forte

A differenza dell'intelligenza artificiale debole in cui il concetto chiave è il problem solving (dunque è presente un nesso diretto tra il problema e la soluzione), l'intelligenza artificiale forte è un approccio profondamente diverso e ambizioso.

Con questo termine si vogliono indicare tutti quei sistemi (indipendentemente dal contesto e dall'obiettivo che viene attribuito) in grado di sviluppare una conoscenza autonoma che non esige in ogni momento di replicare i processi di pensiero tipici dell'uomo, ma che piuttosto punta a sviluppare un'intelligenza generale efficiente in qualsiasi condizione perciò non limitata da particolari necessità.

Se c'è un problema da risolvere, l'intelligenza artificiale debole prova a ipotizzare in modo razionale cosa avrebbe fatto l'uomo in quella precisa situazione mentre l'intelligenza artificiale forte fa riferimento al ragionamento logico e sfrutta i dati disponibili per produrre la conoscenza del contesto da cui derivano le azioni da applicare.

Riassumendo, l'intelligenza artificiale debole punta dunque ad agire razionalmente e pensare umanamente caso per caso (trovando soltanto la soluzione del problema che ne deriva) mentre l'intelligenza artificiale forte è focalizzata maggiormente sull'agire umanamente e pensare razionalmente in termini generali (provando a trovare la soluzioni di tutti i problemi che ne derivano).

Il gioco degli scacchi è un esempio illuminante e piuttosto pratico. La AI debole si limiterebbe a sconfiggere un avversario specifico, analizzando tutte le partite (e le relative mosse) a cui ha preso parte fino ad allora individuando le contromisure al

suo gioco. Invece, la AI forte partirebbe dalla comprensione delle regole del gioco e si allenerebbe in modo costante per diventare sempre più esperta, per battere qualsiasi avversario (e non uno in particolare) diventando così il giocatore di scacchi più forte in assoluto.

Chiarita questa sostanziale differenza tra i due approcci, è ovvio che l'intelligenza artificiale forte presenta (come già anticipato in precedenza) un approccio decisamente più pretenzioso rispetto alla AI debole, e questo nel concreto si traduce in un maggior dispendio di risorse nell'uso delle varie applicazioni.

3.4.5 Funzionamento dell'AI

Per portare a termine le operazioni che è chiamata ad eseguire l'AI, 4 diversi livelli funzionali stanno alla base del suo corretto funzionamento e processo (86):

- Comprensione: l'intelligenza artificiale può riconoscere testi, immagini, video e audio per elaborare determinate informazioni sulla base di una richiesta specifica. Ciò è possibile grazie alla capacità di imparare e simulare la correlazione tra i dati e gli eventi;

- Ragionamento: grazie all'utilizzo di algoritmi matematici appositamente programmati, per l'intelligenza artificiale è possibile collegare autonomamente e logicamente i dati raccolti;

- Apprendimento automatico: i sistemi di Machine Learning (con lo scopo di svolgere funzioni specifiche) sfruttano particolari tecniche per apprendere da un contesto preciso (input) per poi restituire un risultato (output) corretto;

- Interazione: un esempio piuttosto tipico è dato dal Natural Language Processing o NLP, insieme di tecnologie dell'AI che utilizza il linguaggio naturale.

  Come avviene nel caso dei chatbot più innovativi, esso permette di dare origine ad una relazione verbale tra la macchina e l'uomo attraverso sistemi Human Machine Interaction o HMI.

3.4.6 Machine Learning

Il Machine Learning indica un sistema di apprendimento automatico basato sull'AI in grado di assimilare molti dati (input) per allenare una macchina che diventa col passare del tempo sempre più esperta nello svolgere un compito (output) autonomamente, ovvero senza essere stata programmata in anticipo per eseguirlo.

Ciò che contraddistingue un sistema di Machine Learning dal resto è il fatto che esso riesce (senza nessun tipo di ostacolo o impedimento) a generare autonomamente le dovute simulazioni, sfruttando la sua predisposizione all'apprendimento, allo sbaglio e al progressivo e costante miglioraramento dai propri errori fino a diventare sempre più abile.

Un sistema di Machine Learning è piuttosto eterogeneo e fa riferimento a 3 classi principali di algoritmi (86):

1. Con supervisione didattica: il sistema apprende attraverso una correlazione tra input e output da cui impara come prendere una decisione;

2. Senza supervisione didattica: l'apprendimento avviene tramite l'analisi dei risultati senza una relazione diretta tra input e output ma soffermandosi esclusivamente sulla base di output che permettono di mappare i risultati di particolari decisioni, nella stessa situazione in cui i sistemi di ML sono chiamati a fornire e garantire soluzioni;

3. Con rinforzo: il reinforcement learning è un metodo di apprendimento incentrato sul merito, poichè l'intelligenza artificiale viene ricompensata esclusivamente nel momento in cui nelle sue valutazioni consegue un risultato in linea con le aspettative. Grazie all'abilità nel distinguere una decisione corretta da una errata, il reinforcement learning permette di migliorare e perfezionare il training di un sistema machine learning.

Figura 3.7: Machine Learning
(33)

3.4.7 Deep Learning e Reti Neurali

Ritornando al paragone tra l'intelligenza artificiale debole e l'intelligenza artificiale forte, il Machine Learning è uno strumento caratteristico della prima mentre il Deep Learning (che consiste sempre in modelli di apprendimento ispirati al funzionamento del cervello umano) è una tecnica di apprendimento tipica della seconda.

Ciò che distingue quest'ultimo metodo dal machine learning è il fatto che non fa fortemente affidamento alla relazione tra input e output ma piuttosto usufruisce degli input per arrivare a emulare il comportamento del cervello umano.

Il voler riprodurre questo tipo di meccanismo è un qualcosa di affascinante e per far questo molte realtà informatiche prendono spunto dalla struttura di diversi modelli biologici realmente esistenti in natura. Per raggiungere un contesto analogo alle connessioni neurali del cervello umano, il Deep Learning fa riferimento alle cosiddette reti neurali profonde.

Le reti neurali permettono sostanzialmente di riprodurre azioni complesse caratteristiche della mente umana (vedere, parlare, sentire e pensare) tramite l'impiego di neuroni artificiali.

Il termine "profonde" indica reti contraddistinte da molteplici strati di calcolo che fanno riferimento ad una quantià enorme di livelli, tale da esigere uno sforzo di calcolo immenso.

3.5 Intelligenza Artificiale Generativa

3.5.1 Generative AI

L'Intelligenza Artificiale Generativa indica gli algoritmi che possono essere adoperati per generare molteplici contenuti tra cui testo, immagini, codice, audio e video. Il software di intelligenza artificiale generativa parte dalle richieste (prompt) formulate in linguaggio naturale dall'utente (umano o software) e in un secondo momento produce immagini da immagini (Image to Image), immagini da testi (Text to Image), testi da immagini (Image to Text) o testi da testi (Text to Text).

Per istruire e allenare questi algoritmi vengono adoperati dei mix di dati ai fini di generare degli output. Quest'ultimi non sempre rispettano le norme di violazione della proprietà intellettuale dunque potrebbero risultare inopportuni e inadatti.

Tuttavia, l'AI Generativa è in grado di portare risultati apprezzabili sia se dispone di una sorveglianza umana costante e persistente sia se la richiesta dell'utente (i cui feedback possono portare a progressi e sviluppi) risulta pertinente e appropriata.

Attualmente l'intelligenza artificiale generativa è più indicata e adeguata per l'elaborazione di contenuti standard come email, CV o manuali tecnici. Il suo grosso potenziale è però innegabile e viene sfruttato dalle organizzazioni IT e software, che sono in grado di trarre profitto da questi sistemi per produrre in breve tempo una vasta gamma di codici, immagini, testi e sostegno adeguato alla generazione di altri prodotti.

3.5.2 Funzionamento della generative AI

Prima di introdurre il meccanismo che sta alla base della generative AI, è necessario fare un passo indietro e ricordare quali tipologie di apprendimento sfrutta il ML. Quest'ultime possono essere suddivise, sostanzialmente, in apprendimento supervisionato e non supervisionato.

Per cogliere immediatamente la differenza attraverso un esempio pratico, si supponga di avere come insieme di input svariate immagini di animali (con le informazioni relative alle loro caratteristiche).

Nel primo caso i dati sono etichettati, ciò vuol dire che oltre agli input vengono procurati anche gli output (quindi in questo caso il nome dell'animale corrispondente). In questo modo, l'algoritmo sarà in grado di individuare uno schema che lega i dati in ingresso con quelli in uscita e in un secondo momento riuscirà dunque a identificare il singolo animale. Nel secondo caso, invece, i dati non sono etichettati perciò si hanno a disposizione esclusivamente gli input.

In base alle caratteristiche che accomunano maggiormente gli animali, il risultato (ottenuto dall'esecuzione di un sistema di ML su una base dati) sfrutterà un algoritmo adeguato per imparare in modo autonomo a suddividere gli animali in categorie differenti. (88)

Anzichè riuscire a distinguere o classificare la foto di un determinato animale, attraverso una ricombinazione dei dati sfruttati per allenare l'algoritmo, la generative AI prevede di riprodurre una descrizione testuale o un'immagine. Si consideri però che il volume di dati necessario per allenare gli algoritmi è immenso, ben al di là della capacità elaborativa del cervello umano.

Inoltre, i risultati generati possiedono degli elementi casuali. Ciò implica il fatto di avere una moltitudine illimitata di risposte e ciò le fa apparire ancora più creative e fuori dagli schemi.

3.5.3 Vantaggi e sfide

La generative AI può essere utilizzata in pronta consegna così com'è oppure essere allenata introducendo dei propri dati e modelli da cui il software impara. Tra le principali motivazioni che spingono le aziende al suo impiego rientrano le seguenti (74):

1. Ottimizzazione dei processi: può essere utilizzata per migliorare i processi aziendali, come la pianificazione della produzione e distribuzione;

2. Creazione di nuovi prodotti: può essere utilizzata per elaborare nuovi design di prodotto;

3. Miglioramento della Customer Experience: può essere impiegata nella generazione di contenuti personalizzati per i clienti, come raccomandazioni di prodotto o risposte automatizzate ai messaggi dei clienti;

4. Analisi dei dati: può essere impiegata per elaborare enormi volumi di dati e produrre insight che possono dare sostegno alle varie organizzazioni nel prendere decisioni aggiornate;

5. Riduzione dei costi: può assistere le aziende a limitare i costi automatizzando alcuni processi manuali.

L'introduzione e la diffusione esponenziale di una tecnologia così travolgente porta con sé non solo vantaggi ma anche dubbi e limitazioni (70):

1. Qualità dei dati: dati incoerenti o incompleti possono condizionare le performance dei modelli di generative AI, dunque influenzare gli algoritmi necessari per produrre approfondimenti e previsioni di una certa cura e bontà;

2. Privacy dei dati: per riscuotere l'approvazione opportuna da parte dei consumatori, è necessario assicurare che i dati vengano raggruppati e conservati in maniera affidabile;

3. Costi di implementazione: soprattutto le aziende di piccole e media misura potrebbero aver necessità di maggiori risorse per l'implementazione della generative AI, la quale può risultare parecchio onerosa e pretendere degli investimenti considerevoli in termini di personale, infrastrutture e tecnologia;

4. Bias: possono manifestarsi previsioni (e suggerimenti) imprecise o inesatte qualora gli algoritmi venissero allenati su dati incompleti e/o incoerenti; ciò porterebbe ad un'effettiva distorsione;

5. Conformità normativa: nell'usare questa tecnologia bisogna sempre avere riguardo per le diverse normative sulla privacy e sulla sicurezza dei dati (CCPA per gli USA e GDPR per l'UE);

6. Mancanza di esperienza: carenza e scarsità di data scientist, specialisti di intelligenza artificiale e professionisti IT, ovvero il personale tecnico e competente abile nella gestione e creazione di sistemi che sfruttano questo tipo di tecnologia.

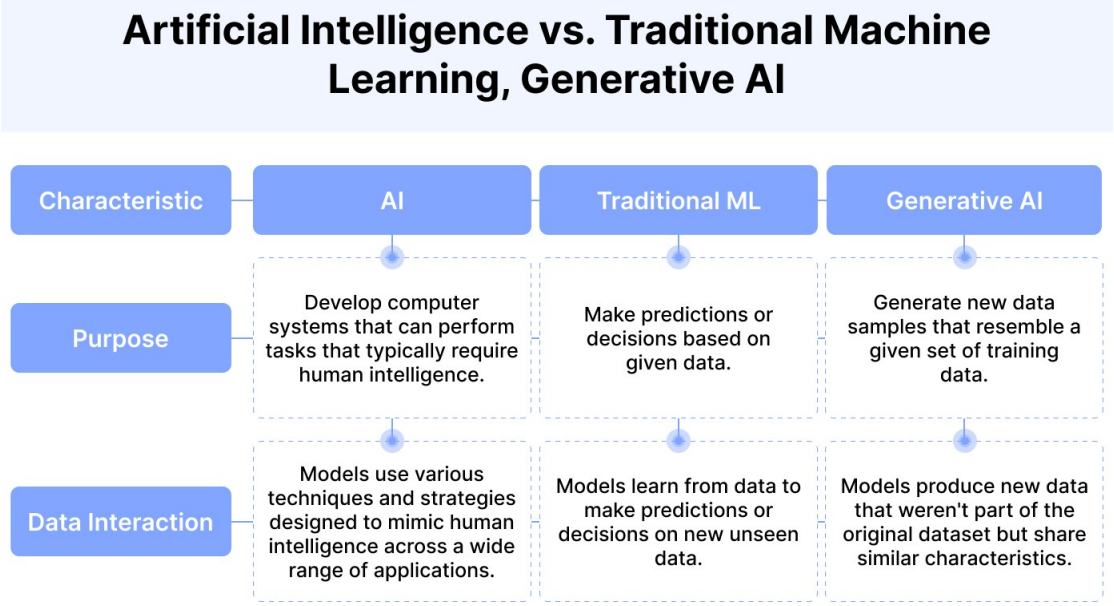


Figura 3.8: Intelligenza Artificiale Generativa (31)

3.6 Predictive Analytics

3.6.1 Analisi predittiva: definizione e funzionamento

L'analisi predittiva (o Predictive Analytics) è un processo/metodo scientifico moderno e innovativo di BI (adattato perlopiù a problemi di business) che, partendo dalla comprensione e dalla formalizzazione del problema, passa tramite la preparazione dei dati e l'analisi esploratoria per poi arrivare a realizzare e mettere in pratica il modello predittivo.

Per far ciò vengono impiegati dati, algoritmi statistici e tecniche di AI e ML per determinare e valutare in maniera ottimale l'attendibilità di risultati futuri facendo riferimento ai dati storici. Questo strumento (estrapolando insight dai Big Data) permette di individuare trend/schemi e acquisisce anticipazioni e stime adeguate sull'intero sviluppo del processo, dimostrandosi sempre più essenziale per le varie organizzazioni. (73)

Con lo scopo di incrementare l'efficienza, la Predictive Analytics viene impiegata nelle aziende per una vasta gamma di applicazioni. Nel marketing si può sapere

con anticipo se un particolare cliente acquisterà o meno una precisa categoria merceologica, dunque possono essere impiegati dei modelli predittivi per realizzare sistemi di raccomandazione sul web, testandone l'efficacia e mettendo in pratica un perfezionamento duraturo e costante.

Nel retail, ML e AI sono impiegati per prevedere l'andamento della domanda, migliorare l'assortimento e prevenire le rotture di stock. Questa tecnologia viene anche adoperata nelle istituzioni finanziarie (ad esempio le banche) per individuare le frodi e venire a capo dei problemi di conformità.

Come anticipato in precedenza, l'analisi predittiva fa uso di algoritmi che fanno riferimento ai dati storici, i quali necessitano di una precisa preparazione (pulizia, filtraggio, categorizzazione) prima di poter godere di una certa qualità e adeguatezza.

Essi sono raccolti tramite vari strumenti e canali e vengono implementati con modelli di analisi che possono essere statistici, basati sul cosiddetto machine learning superficiale (alberi decisionali, random forest, gradient boosting e così via) oppure possono essere algoritmi più all'avanguardia e innovativi, come ad esempio i network bayesiani o il deep learning.

Tuttavia, la Predictive Analytics ha fatto la sua prima apparizione esclusivamente nel momento in cui le varie imprese e organizzazioni hanno potuto usufruire (mediante velocità di elaborazione elevate) di tecnologie moderne evolute come infrastrutture di data mining, software predittivi e altri algoritmi di una certa complessità.

Grazie a questi nuovi strumenti, le aziende sono in grado di poter accedere a immense quantità di dati strutturati, semi-strutturati e non strutturati ed estrarre valore e conoscenza (insight) a sostegno delle loro attività e decisioni strategiche.

3.6.2 Motivi per servirsi dell'analisi predittiva

Grazie all'utilizzo di tecniche statistiche, l'analisi predittiva è in grado di far comunicare set di dati e algoritmi personalizzati per evidenziare possibili scostamenti o anomalie di modelli precedenti, raccomandando cambiamenti o individuando errori in esecuzione. In tal maniera si anticipano esiti futuri come cambiamenti sui mercati, scelte dei consumatori e così via.

Evidentemente non sarà una previsione corretta e impeccabile, ma quel che è certo è che viene consentito di prevedere prima i problemi e di rispondere alle nuove tendenze per agire di conseguenza, realizzando piani d'azione che garantiscono una maggior efficienza.

In tempi brevi, il vantaggio più immediato si verifica (misurandone l'effetto in modo tempestivo) nella predizione applicata a processi operativi. Benefici a medio e lungo termine si ricavano solitamente integrando dati interni ed esterni, i quali consentono di produrre analisi e simulazioni predittive a sostegno di monitoraggio e adattamenti della strategia.

3.6.3 Modelli di analisi predittiva

Per permettere agli utenti di elaborare i dati passati e presenti in intuizioni concretizzabili (producendo risultati favorevoli a lungo termine), è necessario introdurre alcuni tipi di modelli che stanno alla base dell'analisi predittiva (66):

- Customer Lifetime Model: identifica i clienti che hanno maggiori possibilità di investire in servizi e prodotti;
- Customer Segmentation Model: classifica i clienti in base a caratteristiche e comportamenti di acquisto affini;
- Predictive Manteinance Model: consente di prevedere eventuali rotture di apparecchiature fondamentali;
- Quality Assurance Model: identifica e previene i difetti per evitare problemi e costi addizionali nel momento in cui si forniscono servizi o prodotti ai clienti.

3.6.4 Algoritmi di analisi predittiva

Abitualmente, gli algoritmi di analisi predittiva sono ideati e sviluppati per venire a capo di un determinato problema aziendale o una serie di problemi, ai fini di ottimizzare un algoritmo esistente o per procurare un tipo di capacità unica. Questi algoritmi possono essere di varia natura: di tipo statistico, di data-mining o di ML.

A seconda dei casi è più utile sfruttare una tipologia piuttosto che l'altra: in genere, un algoritmo di regressione viene selezionato per prevedere l'esito di molti

eventi temporali o per generare un sistema di credit scoring (sistema automatizzato adottato dalle banche e dagli intermediari finanziari per valutare le richieste di finanziamento della clientela).

Invece, per ottimizzare la fidelizzazione dei clienti o per potenziare un sistema di recommendation (suggerimenti verso il consumatore per guidarlo verso la finalizzazione di un acquisto) si sfruttano maggiormente gli algoritmi di classificazione. Ad esempio, gli algoritmi di clustering sono idonei e appropriati alla segmentazione dei clienti, al rilevamento delle community e ad altre attività di tipo social.

Capitolo 4

Customer Analytics in ambito Retail

L'obiettivo del capitolo è descrivere nel dettaglio tutto ciò che è affine all'analisi dei dati del mondo Retail. Con questo termine ci si riferisce a tutte quelle attività che sono associate alla vendita di prodotti/servizi da parte di un'azienda (detta retailer) direttamente al consumatore finale.

Nella definizione rientrano sostanzialmente la Grande Distribuzione Organizzata o GDO (ovvero i supermercati), la Grande Distribuzione Specializzata o GDS (che a differenza delle GDO operano in un unico settore), il fashion, il beauty e così via.

Il motivo per il quale si andrà ad analizzare il mondo Retail sta nel fatto che quest'ultimo è uno dei mercati che può trarre maggiormente vantaggio dalla veloce e profonda evoluzione delle tecnologie e metodologie di una Data Platform.

Il Retail porta con sè tematiche classiche comuni a tutte le aziende (analisi delle vendite, della logistica, della supply chain, etc) ma ne ha altre peculiari e molto più interessanti tra cui soprattutto la Customer Analytics.

Il termine Customer Analytics indica l'insieme di processi e tecnologie che permette alle aziende di raccogliere e apprendere la conoscenza del cliente. Queste attività consentiranno di adottare decisioni chiave che porteranno all'avanzamento e al suggerimento di offerte tempestive e pertinenti verso i propri clienti. (67)

Ci sono molteplici strumenti che servono a fare Customer Analytics. Alcuni di questi verranno descritti in questo capitolo e sono la UCV, la targetizzazione dei

clienti a supporto di una campagna marketing, il volantino personalizzato, il Churn Rate, il CLV e il TtNP.

4.1 UCV (Unique Customer View)

La UCV (Unique Customer View) consiste in una rappresentazione aggregata, coerente e olistica che permette di archiviare ogni informazione sui clienti in un luogo univoco centralizzato. Tutto ciò consente di avere una panoramica puntuale sugli utenti stessi, in cui è possibile visionare ogni loro tipo di movimento e interazione. (84)

Inoltre, la UCV contiene una serie di indicatori chiave di performance (Key Performance Indicators o KPI), i quali fungono da etichette con lo scopo di andare poi a descrivere il cliente. Essi variano a seconda degli obiettivi e progetti dell'azienda ma possono essere generalizzabili a prescindere dal tipo di cliente che si ha di fronte.

4.1.1 KPI standard della UCV

Affrontando nello specifico la tematica, è necessario andare a discutere tutti quelli che possono essere i KPI principali che caratterizzano la struttura classica di una UCV per il mercato Retail.

Tra i più importanti ci saranno sicuramente quelli inerenti all'anagrafica dei clienti/articoli/negozi, alle transazioni dello scontrinato, ai comportamenti d'acquisto nei punti vendita o online, alle promozioni/sconti, alle tematiche di loyalty e così via.

Alcuni KPI (in quanto campi sensibili) dovranno essere protetti attraverso uno strumento software che individua e previene l'uso non autorizzato e la trasmissione di informazioni riservate. Tipicamente, questa tecnologia di sicurezza viene chiamata Data Loss Prevention (d'ora in poi per semplicità DLP).

Ecco ora la lista dei KPI standard che deve avere necessariamente una UCV:

1. KPI inerenti all'anagrafica clienti
    - Data partizione: è l'indicazione del giorno che funge da campo di partizionamento. Si mette una data di inizio e una di fine con lo scopo di

limitare la validità del record. Questo deriva dal fatto che la UCV viene generalmente storicizzata. Esistono almeno due metodi di storicizzazione.

Il primo consiste nello storico a snapshot, nel quale si hanno delle istantanee giornaliere/settimanali che vengono archiviate all'interno di una partizione. In caso di eventuali crash del server, si perderanno esclusivamente i record relativi all'intervallo di tempo trascorso dall'ultima istantanea rilevata.

Attraverso una coda ciclica nella storicizzazione è possibile garantire il rispetto di alcuni criteri del GDPR (General Data Protection Regulation) e limitare la profondità storica di alcuni dati.

Il secondo metodo di storicizzazione è invece l'SCD2 o Slowly Changing Dimension di tipo 2, uno storico che viene fatto normalmente su tabelle degli stati ovvero quelle caratteristiche che cambiano in tempi non brevi, come ad esempio l'anagrafica.

L'obiettivo è tenere traccia della cronologia degli aggiornamenti ai record della dimensione, mantendendo sia il vecchio record che quello nuovo (aggiunto in una nuova riga). Ecco un esempio illustrativo:

customer_id	customer_name	email	start_date	end_date
1	John Doe	john.doe@example.com	2022-01-01	null
2	Jane Smith	jane.smith@example.com	2022-02-01	null

Figura 4.1: SCD2 (step 0)
(71)

Si supponga di essere nella situazione riportata in figura e che il 15 marzo 2022 John Doe aggiorni il suo indirizzo email. Ciò che bisogna fare è creare un nuovo record per lo stesso cliente con l'indirizzo e-mail aggiornato.

La end data del vecchio record viene impostata sul giorno dell'aggiornamento (contrassegnandolo come "scaduto") mentre la start date del nuovo record è impostata sul giorno dell'aggiornamento (indicando la modifica).

customer_id	customer_name	email	start_date	end_date
1	John Doe	john.doe@example.com	2022-01-01	2022-03-15
1	John Doe	johndoe@example.com	2022-03-15	null
2	Jane Smith	jane.smith@example.com	2022-02-01	null

Figura 4.2: SCD2 (step 1)
(71)

- Numero cliente: numero identificativo del cliente che funge solitamente da chiave primaria della UCV. A seconda dei casi, può essere tenuto in chiaro oppure cifrato per questioni di privacy.
- Codice fiscale: codice fiscale del cliente, che in quanto campo sensibile deve essere protetto attraverso lo strumento software di sicurezza DLP.
- Nome, cognome e indirizzi relativi a domicilio/residenza: nome, cognome e indirizzo del cliente. In maniera del tutto simile a quanto accade per il codice fiscale, anche questi campi evidentemente sensibili devono essere codificati attraverso lo strumento DLP.
- Data di nascita/età: data di nascita del cliente, dalla quale è possibile ricavare l'età. Il problema che sorge in questo caso è il fatto che il dato arriva cifrato poichè la data di nascita viene ritenuta un attributo sensibile (si potrebbe ricostruire il codice fiscale).

  Quindi l'approccio corretto consiste nel tenere il campo "anno di nascita" in chiaro e il campo "data di nascita" cifrato.
- Codice negozio radicamento: codice identificativo del negozio di radicamento del cliente. Solitamente, si hanno relazioni 1 a N in quanto un cliente può avere più punti vendita associati.

  Avendo in UCV una sola riga per rappresentare l'utente, bisogna associare il negozio prevalente in base a (tendenzialmente) dove apre il suo conto, dove va più spesso oppure dove spende più denaro.

  Oltre al codice può essere memorizzata la posizione geografica e un attributo "descrizione".

- Canale: canale del negozio di radicamento. È abbastanza ricorrente trovare questo indicatore nei supermercati, ipermercati, minimarket e negozi di prossimità.
- Data inizio relazione: data di inizio della relazione tra il cliente e il brand. Essa può essere determinata in due modi differenti. Si potrebbe considerare la data in cui il cliente fa la tessera oppure si potrebbe considerare la data in cui il cliente fa la prima spesa.
- Anzianità cliente: distanza (solitamente in mesi) tra la data di elaborazione della Unique Customer View e la data di inizio relazione tra cliente e brand.

  In base ad essa, il cliente viene categorizzato sulla base di diversi segmenti e quindi viene associato a etichette specifiche (un esempio di segmentazione del cliente potrebbe "nuovo", "recente", "consolidato" e "fidelizzato").
- Email e cellulare: contatti email e di cellulare del cliente che vengono messi a disposizione del brand. Come già accadeva per il numero cliente, il codice fiscale, nome, cognome e data di nascita, anche in questo caso tali contatti devono essere cifrati.
- Opt in 1, opt in 2, opt in 3: termini che fanno riferimento ai consensi di marketing che il cliente decide di fornire o meno. Nell'ordine sono rispettivamente "consenso alla profilazione", "marketing diretto" e "comunicazione dei dati a soggetti terzi".

  Sono tutte molto importanti, soprattutto la prima, perchè se un cliente non vuole essere profilato a cascata tutto ciò comporta una serie di conseguenze.

  A seconda del contesto, se un cliente ha negato la sua profilazione possono verificarsi due scenari. Il primo prevede inizialmente l'inserimento del cliente in UCV ma in seguito i suoi dati non vengono forniti al software di marketing automation che si occupa di inviare le comunicazioni. Il secondo prevede invece che il cliente non finisca neanche in UCV e quindi sparirà fin da subito.

  Questo dipende molto dal DPO (Data Protection Officer) ovvero la figura interna/esterna che si occupa di garantire le regole di protezione dei dati.

Tutti i non profilabili diventano un'entità anonima che viene aggregata, dunque i relativi indicatori vengono calcolati in forma aggregata.

- Mezzo contatto preferito: questo indicatore definisce il touchpoint più utilizzato da parte del cliente. Esso può corrispondere al cellulare, alla email, ad entrambi o a nessuno.

2. KPI inerenti all'anagrafica tessere

   In alcune GDO potrebbe sussistere una relazione 1:N per quanto riguarda le tessere per cliente. Avendo solo una riga per rappresentare queste informazioni, si sceglie una carta che è prioritaria rispetto alle altre.

   La tipologia di carta determina il criterio di scelta della stessa, che dovrà poi essere cifrata per i soliti motivi di privacy. In altre GDO, la situazione è più semplice in quanto la tessera è nominale dunque univoca per ogni cliente.

3. KPI inerenti al comportamento dei clienti online

   In modo analogo a quanto detto poc'anzi, gli stessi indicatori possono essere estratti dal comportamento dei clienti online:

   - Data creazione e data modifica profilo: rispettivamente data di creazione del profilo online del cliente e data dell'ultima modifica effettuata.
   - Anzianità: questo indicatore corrisponde al lasso di tempo che intercorre dalla data di iscrizione del cliente fino alla data odierna. Una parte dei clienti fidelizzati tende ovviamente ad avere un alto valore di anzianità, altrimenti il loro rapporto con il brand sarebbe terminato con grande anticipo.
   - Login web/app: indicatore che monitora l'attività del cliente nel mondo online, come ad esempio le interazioni con il portale, il sito web o l'app mobile del retailer.
   - Recenza (online): tempo che è trascorso dall'ultima visita online del cliente. Questo lasso di tempo si misura solitamente in termini di giorni.

4. KPI inerenti alle transazioni/scontrinato

   Fonte dati più voluminosa che arriva solitamente in maniera sincrona. Ciò significa che il software di cassa raccoglie localmente i dati e gli scontrini, e a fine giornata (quando chiudono i punti vendita) invia lo scontrinato del giorno.

   A volte si perdono alcuni dati dunque quest'ultimi devono essere reinviati nei giorni successivi per far fronte ad eventuali modifiche o integrazioni, in modo tale da rendere il tutto più stabile.

   - Numero carta più utilizzata: numero identificativo della carta più utilizzata da parte del cliente. Questo campo è utile nei casi di GDO che prevedono l'impiego potenziale di più carte da parte di un cliente e, al netto della priorità che viene espressa, si può decidere di usare alla cassa un tipo di carta piuttosto che l'altro.

     Più utilizzata significa che si guarda il maggior numero di scontrini registrati associati, e una volta fatto questo si cifra il campo (come già accadeva per alcuni dei precedenti).

   - Frequenza negozio radicamento: nella finestra temporale stabilita, corrisponde al numero di visite (in giorni) che il cliente effettua nel suo negozio di radicamento (dove il cliente fa la tessera).

   - Canale d'acquisto prevalente: identifica semplicemente il canale prevalente del cliente a livello di spesa effettuata.

   - Spesa per canale: questo indicatore mostra la spesa effettuata da parte del cliente per ogni specifico canale. Può tornare utile nel momento in cui si vuole sapere su quale canale il cliente preferisce acquistare.

   - Id negozio spesa/frequenza prevalente: codice identificativo del negozio in cui il cliente ha avuto un indice di spesa/frequenza maggiore. Per poterlo individuare e recuperare, si stila una graduatoria di tutti i negozi in cui è stato il cliente e si ordinano per spesa/frequenza decrescente selezionando il codice del primo negozio.

     Tra spesa e frequenza c'è ovviamente una correlazione piuttosto alta dunque tendenzialmente una implica l'altra e viceversa.

- Giorno acquisto e fascia oraria prevalente: questi indicatori mostrano rispettivamente il giorno d'acquisto e la fascia oraria prevalente in cui si presenta in negozio un cliente.

  Anche in questo caso esiste un concetto di ranking, ovvero a seguito di un ordinamento decrescente dei giorni della settimana e delle fasce orarie, si seleziona la prima voce in entrambe le classifiche.

- Recenza: misura che viene calcolata contando il numero di giorni trascorsi dall'ultima visita del cliente al punto vendita. Dunque indica in qualche modo quanto spesso l'utente si reca presso lo stesso punto vendita.

- Frequenza: misura che conta o il numero di scontrini che il cliente fa per unità di tempo o il numero di giorni distinti di sue visite al negozio.

  Entrambe le definizioni sono corrette e il motivo per il quale si considerano entrambe (nonostante siano altamente correlate) è dovuto al fatto che in questo modo è possibile escludere i clienti che fanno più spese in un giorno.

- Intertempo medio acquisti: questo indicatore esprime il lasso di tempo medio che intercorre tra due acquisti successivi di un cliente. Questa è la prima misura che suggerisce come il cliente potenzialmente si stia disaffezionando.

  È all'incirca il reciproco della frequenza (ogni quanto viene visto un cliente in negozio) ed è interessante perchè è la base per andarci a costruire sopra degli algoritmi.

- Valore monetario: corrisponde alla spesa totale effettuata dal cliente durante il periodo stabilito. Se tale valore viene spacchettato a livello di reparto, allora il tutto assume maggior significato. In questo modo sarà infatti più rapida e semplice l'individuazione dei reparti che hanno originato le marginalità più elevate.

- Scontrino medio: corrisponde al rapporto tra la spesa totale effettuata dal cliente e il relativo numero di scontrini registrati. Per lo stesso motivo del valore monetario, questo indicatore è molto importante ed è da tenere in considerazione in quanto se viene applicato per reparti assume una

maggiore utilità. Infatti, sarà possibile trovare il settore che produce i profitti più alti.

- Recenza meno intertempo medio ultimi acquisti: differenza tra la recenza e il lasso di tempo medio che intercorre tra due acquisti successivi del cliente in riferimento all'ultimo periodo.

  Se questa misura assume un valore negativo non sorge nessun tipo di problema, in caso contrario bisogna iniziare a preoccuparsi (è necessario fare comunicazioni ad hoc al cliente).

  È uno degli indicatori più utili in quanto tutti i vari motori di raccomandazione utilizzano delle variabili preaggregate per alimentare un algoritmo di questo tipo.

- Spesa/frequenza media mensile rolling: questo indicatore va a definire il valore cliente medio (RFM o recenza - frequenza - valore monetario) su finestra rolling (concetto di media mobile). Modo piuttosto semplice per attribuire un valore al cliente.

  Solitamente si costruisce una griglia 3x3 (un asse è la spesa media mensile, l'altro è la frequenza media mensile) che va a suddividere i cluster cliente rolling (bronze, silver o gold). Le stesse classificazioni le possiamo immaginare non rolling (periodi fissi a calendario).

- Valore cliente: indicatore che cerca di capire che tipo di cliente si sta valutando, quindi essenzialmente mostra al brand il relativo potenziale di spesa. Ci sono situazioni in cui si preferisce utilizzare la somma o la spesa media mensile, altre in cui si usa lo scontrino medio (che è incorrelato rispetto alla frequenza).

- Reparto misure: array che contiene una serie di misure calcolate sul reparto, che è uno dei livelli della classificazione marketing. L'alternativa è creare un campo per ogni misura ma l'impiego di un array permette una certa dinamicità nell'ambiente di un database relazionale.

  Ad esempio, è possibile calcolare la frequenza/spesa per reparto, misura che indica rispettivamente il numero di visite che il cliente fa al reparto e la spesa totale effettuata per ciascuno a livello di prodotti.

5. KPI inerenti alle promozioni e agli sconti

   Per quanto riguarda le promo e gli sconti, gli indicatori servono a valutare quanto il cliente acquista in promozione.

   - Percentuale promo valore: indica il valore finale scontato a seguito di una promozione. Se ad esempio un cliente avrebbe dovuto spendere 100 ma con gli sconti ha speso 90, allora la misura assume valore 0.1.
   - Percentuale promo quantità: indica la quantità di prodotti acquistati a seguito di una promozione, quindi è un concetto analogo all'indicatore precedente ma che non ragiona per valore. Per fare un esempio, su un totale di 100 articoli acquistati basta considerare quanti di questi sono in promozione.
   - Prodotti sempre in promozione: termine che indica tutti quei prodotti che risultano in promozione in ogni momento. Ciò porta a distinguere quest'ultimi da quelli che possono essere trovati in promozione solamente in occasioni speciali e quindi in situazioni specifiche.
   - Cassaforte sconto: indica un accumulatore (o appunto una cassaforte) di scontistiche, che vengono applicate di diritto in misura maggiore ai clienti con i valori più alti di spesa.

     Questo è un modo per incentivare sia gli utenti che già possiedono questo tipo di profilo sia gli utenti con valori di spesa più bassi.
   - Valore risparmio: cifra in euro che il cliente ha risparmiato con le diverse promozioni nell'ultimo anno rolling. È la controparte in valore assoluto dello sconto. Invece, l'indicatore "percentuale promo valore" (elencato precedentemente) esprime l'equivalente in percentuale.
   - Valore risparmio year to date: cifra in euro che il cliente ha risparmiato con le diverse promozioni nell'ultimo anno non rolling. Quindi è una declinazione di "valore risparmio" in quanto questa volta si parte col calcolo dal primo gennaio.
   - Score "nome reparto": formalmente indica un percentile di spesa del cliente all'interno di un determinato reparto. È un indicatore numerico che varia

tra 0 e 1 e più lo score relativo ad un cliente è elevato, più quest'ultimo avrà acquistato a livello monetario.

- Indice di Gini: indicatore di varietà che (in presenza di una variabile categorica) dice quanto è uniforme la distribuzione tra le diverse categorie oppure quanto essa è sbilanciata.

  Lo 0 indica la varietà minima mentre l'1 indica la varietà massima. È utile ad esempio a definire la varietà di acquisto di un cliente in tutte le categorie che compongono un reparto merceologico.

6. KPI inerenti al comportamento dei clienti rispetto a programmi di Loyalty ed iniziative commerciali e di marketing

   La fidelizzazione nei normali supermercati si limita a vedere se il cliente fa la raccolta dei punti monitorando tendenzialmente i movimenti che esso esegue (spendendo si maturano dei punti che poi possono essere usati in diversi modi).

   Molto spesso esistono delle fonti dati che descrivono il rapporto tra il cliente e la GDO che va oltre il solo scontrinato. Gli scontrini non costituiscono un dato certificato. Se si possiedono delle informazioni di flusso (fatti) che si succedono nel tempo, una complicazione è avere n fatti associati.

   Ci sono 2 livelli di complessità per rappresentare questo dato: il primo è che ci può essere una relazione 1 a N, quindi siccome la UCV in realtà rappresenta relazioni 1 a 1 (ha come chiave il cliente) bisogna trovare la chiave di aggregazione; il secondo è che questi eventi si succedono nel tempo e in UCV non si gestisce il tempo come dimensione, perciò bisogna convertire tutto in una sola riga.

   Ciò viene fatto scegliendo a priori uno o più orizzonti temporali (1 anno, semestre, quadrimestre, 8 settimane) in cui si va a fare il calcolo delle misure aggregate. Anzichè aggiungere delle dimensioni che descrivono il periodo di tempo all'orizzonte (andrebbero a moltiplicare i record), quello che si può fare è moltiplicare le misure.

   12 mesi è un'ampiezza temporale che permette di avere una visione di ampio orizzonte e consente di annullare gli effetti della stagionalità. Tale fenomeno

è piuttosto tipico nel retail, basti pensare ai supermercati che subiscono fortemente sia picchi che cali. Dunque possono essere confrontati diversi anni rolling.

4.1.2 Misure abilitanti per l'uso di algoritmi

Sfruttando i KPI descritti nel paragrafo precedente, l'idea è costruire delle misure che sono abilitanti rispetto all'uso di algoritmi tipici in quest'ambito. Oltre alla finalità legata al CRM e alla profilazione del cliente (quest'ultima utile per fare monitoraggio però fine a se stessa in quanto dà poche informazioni), l'UCV ha anche quest'altra funzione.

Le cose interessanti per gli analisti dei dati sono due: la prima consiste nell'utilizzare queste informazioni per attivare una comunicazione personalizzata verso il cliente (approccio utile per provare a incrementare la frequenza di spesa, portare delle offerte e così via) mentre la seconda mira ad abilitare la parte vera e propria di algoritmi.

Alcuni di questi possono essere ad esempio il Churn Rate (tasso di abbandono del cliente), il TtNP / Time to Next Purchase (nel contesto GDO si cerca di modellare gli intertempi d'acquisto come variabile non più categorica ma come una variabile numerica similcontinua che corrisponde alla distanza in giorni tra due acquisti successivi del cliente) o il CLV / Customer Lifetime Value (ha l'obiettivo di capire in quale fase della vita si trova il cliente e qual è il suo potenziale di spesa).

Il rapporto con il marchio può essere modellato come delle curve di sopravvivenza, curve decrescenti che misurano la durata del rapporto tra il cliente e la GDO. Il modo in cui si comporta il cliente assegna una di queste curve. Tutto ciò può dare una sorta di stima di quella che è la sua capacità di spesa nel tempo (come se fosse l'integrale di queste curve).

L'UCV può diventare una raccolta dati abilitante per questi algoritmi, perchè tendenzialmente semplifica processi di calcolo che altrimenti si dovrebbero fare a partire da delle tabelle grezze disaggregate che possono diventare molto pesanti.

L'efficacia di una UCV dipende dalla capacità dell'azienda di impiegare tali metriche per adattare e ottimizzare costantemente le proprie strategie di marketing e customer experience.

4.1.3 Sorgenti dati che alimentano la UCV

Per la creazione della UCV sono necessarie sorgenti dati di diversa natura, alcune delle quali sono pressochè standard nei progetti in ambito Retail.

1. CRM (Customer Relationship Management): software che memorizza sia le informazioni di contatto dei clienti (indirizzo e-mail, numero di telefono, profilo sui social media) sia le interazioni del cliente stesso con l'azienda. Può anche registrare dati che indicano ad esempio le preferenze personali dei clienti a livello di comunicazione (se preferiscono via telefono piuttosto che via email).

   Dunque si tratta di un software che ottimizza la gestione di questi rapporti e che di conseguenza dà origine a conversazioni produttive ed efficaci. Una delle aziende che ricopre un ruolo centrale nel contesto CRM è senza dubbio Salesforce. (45)

Figura 4.3: CRM (23)

2. Sistemi Retail: strumenti che vengono utilizzati per la gestione operativa dei punti vendita. In questa categoria sono inclusi anche i sistemi cassa e gli e-commerce, sorgenti in tempo reale che sfruttano strumenti che sono stati configurati ragionando secondo un concetto di coda.

   Il termine coda significa che le informazioni in ascolto su questi canali vengono raccolte in base a dei topic e il sistema in questione le scoda ovvero le mette a disposizione di un sistema di trasformazione del dato.

   Attraverso questa sorgente è possibile analizzare diversi dati tra cui principalmente transazioni dello scontrinato e anagrafiche.

   - "Transazioni dello scontrinato"

Figura 4.4: Transazioni Scontrinato
(36)

Contiene le informazioni sullo scontrinato quindi il dettaglio di ogni transazione. Per la loro gestione, elaborazione (in tempo reale), integrità e sicurezza viene impiegato un insieme di tecniche software che prende il nome di OLTP (On-Line Transaction Processing).

Questo sistema si presta per l'inserimento, l'aggiornamento e l'eliminazione di dati. Dunque i dati delle transazioni online sono la fonte dei dati per OLTP, il quale impiega solo database relazionali. (46)

- "Clienti"
  Anagrafica dei clienti che effettuano spese all'interno della GDO in questione.
- "Articoli"
  Anagrafica degli articoli, la quale contiene informazioni come codice del prodotto, categoria alla quale appartiene e così via.
- "Carte"
  Anagrafica delle carte con le informazioni relative alla tipologia della carta o al numero del cliente.
- "Dim enti"
  Dimensione analitica che descrive gli enti ovvero i punti vendita.

3. ERP (Enterprise Resource Planning): sistema gestionale che integra tutti i processi di business rilevanti di un'azienda (vendite, acquisti, gestione magazzino, contabilità, finanza, etc).

Figura 4.5: ERP
(29)

4. CMS (Content Management System): applicativo web che nel contesto del CRM analitico funge da fonte alimentante in grado di descrivere (tramite i dati raccolti sui device e sui social) i comportamenti dei clienti online.

Figura 4.6: CMS
(22)

5. Loyalty clienti: fonte dati che tra le varie informazioni raccoglie, per esempio, accumulo/fruizione punti (in cassa e su piattaforme online) e partecipazioni ad attività non di vendita promosse dalla GDO (azioni di solidarietà, donazioni, etc).

Per la definizione della UCV, alcune informazioni relative a determinate sorgenti dati risultano essenziali mentre altre possono essere considerate opzionali.

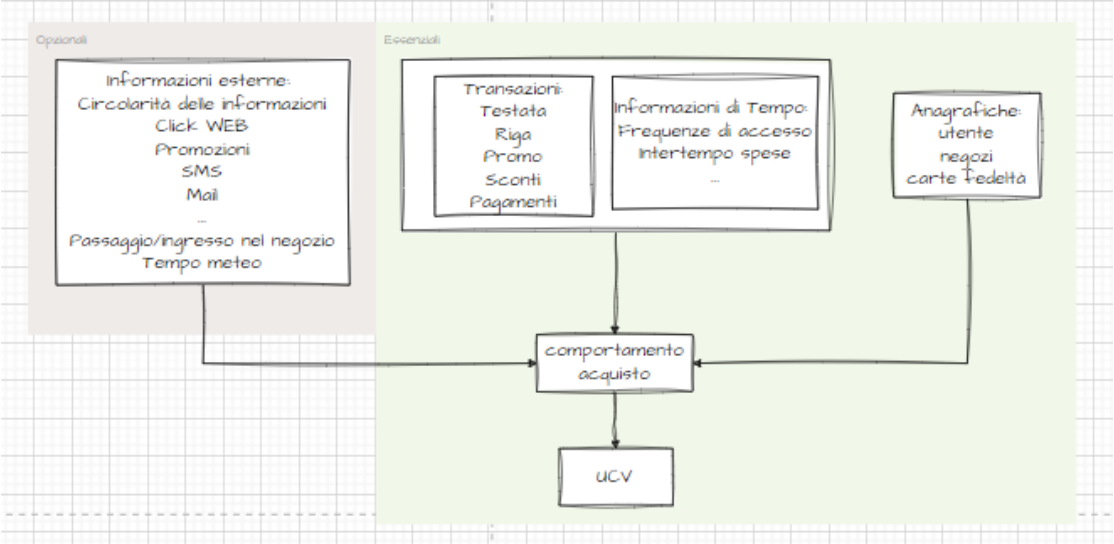


Figura 4.7: Sorgenti
(68)

4.2 Targetizzazione dei clienti a supporto di una campagna marketing

Per qualsiasi analisi mirata ai clienti è fondamentale che essi siano profilati attraverso la carta/tessera fedeltà. Senza quest'ultima non sarebbe possibile svolgere nessun tipo di analisi.

Qui entra in gioco il concetto di UCV perchè una volta che si riescono a definire tutti gli indicatori più utili (monitorandoli e calcolandoli sulla base di un intervallo temporale sufficientemente lungo), è possibile ottenere una vista già pronta per poter avviare il processo di targetizzazione poichè si riesce attraverso dei filtri a individuare i clienti a cui si vuole mirare.

In un secondo momento sarà possibile attivare la personalizzazione dei messaggi di comunicazione grazie proprio a questa procedura di segmentazione in target list (paragrafo successivo).

Il punto da cui un'azienda deve partire consiste nell'impiegare i dati disponibili per conoscere meglio il pubblico di riferimento, soprattutto per individuare il segmento di clienti che presumibilmente genererà il valore superiore nel tempo (e quindi i migliori risultati da un punto di vista prettamente economico).

4.2.1 Approccio nell'individuazione del giusto target

La targetizzazione dei clienti è un processo che richiede tempo e precisione. In tal caso, è necessario adottare delle best practices o un approccio che consenta alle aziende di ottenere i migliori risultati al fine di individuare i giusti target. Sono previste diverse fasi, che sono riportate qui di seguito (63) - (60):

1. Fase di analisi

   - Analisi del pubblico attuale: bisogna capire e comprendere chi sono i clienti attuali del brand. Per far ciò è necessario raccogliere su di loro dati demografici, comportamentali e geografici, con i quali sarà possibile avere un'idea di chi sta già rispondendo in modo positivo all'azienda.
   - Definizione delle buyer personas: le buyer personas descrivono le categorie di utenti che sono più propensi a comprare determinati prodotti o servizi. Il termine indica dunque tutti i profili dettagliati di coloro che possono essere definiti clienti ideali. Le buyer personas includono diverse informazioni tra cui età, genere, occupazione, interessi, obiettivi ma anche problemi.
   - Ricerca di mercato: per andare alla ricerca dei clienti che sono idealmente da raggiungere, quest'analisi risulta strettamente necessaria ai fini di individuare opportunità nel mercato e tendenze attuali nel settore di competenza.

2. Fase di segmentazione

   A seconda del contesto e del progetto che sta affrontando l'azienda, l'attività di segmentazione può essere fatta in un modo piuttosto che in un altro. Esistono diversi tipi di segmentazione, tra cui i principali risultano i seguenti:

   - Segmentazione demografica
     Approccio tra i più semplici poichè richiede informazioni facilmente reperibili (per esempio da dati di terzi come i censimenti) quali l'età, l'istruzione, il reddito familiare, lo stato civile, le dimensioni del nucleo familiare, l'etnia, il genere, l'occupazione e la nazionalità. Inoltre, è uno degli approcci più impiegati poichè tutti questi fattori influiscono spesso sul cliente.

- Segmentazione geografica
  In questo caso si dà origine a gruppi differenti di clienti sulla base di confini geografici. Questo tipo di segmentazione può essere considerata un sottoinsieme di quella demografica e può facilitare il compito di un'azienda nello stabilire dove pubblicizzare (per poi vendere) i propri prodotti, nonché cercare il luogo in cui espandersi.

  Attraverso sondaggi, ricerche di mercato di terzi e dati operativi (come ad esempio gli indirizzi IP dei visitatori del sito web), è possibile dunque esaudire le preferenze, gli interessi e le esigenze dei propri clienti.

- Segmentazione comportamentale d'acquisto
  Questa particolare targetizzazione è basata sui comportamenti d'acquisto e permette ai retailer di seguire un approccio più mirato, in quanto possono concentrarsi su ciò che i clienti vogliono o sono più inclini ad acquistare.

  A differenza degli altri tipi di segmentazione, in questo caso è consigliabile sfruttare i dati e gli insight già a disposizione dell'azienda piuttosto che affidarsi a ricerche di mercato di terzi.

- Segmentazione psicografica
  Questa segmentazione è impiegata maggiormente nei mercati ampi. In tale contesto, i clienti vengono raggruppati per il loro stile di vita, tratti della personalità, valori, opinioni e interessi dunque secondo i loro aspetti psicologici.

  Le informazioni più attendibili e credibili per questo tipo di targetizzazione sono quelle fornite dai diretti interessati attraverso dei sondaggi ad hoc. Infatti, quest'ultimi presentano domande qualitative che permettono ai clienti stessi di condividere le loro informazioni utili.

  Nonostante venga restituita una profilazione molto precisa del singolo cliente, purtroppo tali sondaggi hanno una copertura molto bassa (perchè costano molto), pertanto rischiano di rappresentare una porzione molto ridotta della customer base o comunque un campione non rappresentativo.

3. Fase di valutazione

   - Valutazione dei touchpoint: è necessario individuare i touchpoint (punti vendita fisici, siti web, e-commerce o profili social) che i clienti adoperano maggiormente.
   - Test, raffinamento e ottimizzazione: premettendo che bisogna continuare a raccogliere dati e ad analizzarli (per avere sempre una strategia di targeting che sia dinamica e adattabile ai cambiamenti continui del mercato), le campagne di marketing vanno costantemente monitorate, testate, raffinate ed infine ottimizzate.

     Mediante strumenti analitici sarà così possibile valutare quali campagne funzionano meglio e quali meno, in modo tale che possano essere apportate le dovute modifiche per migliorare l'approccio.
   - Feedback dei clienti: bisogna interrogare i clienti per venire a conoscenza dei pensieri riguardanti le campagne di marketing attuate. Tale feedback risulterà prezioso per affinare ancora di più le target list mediante suggerimenti utili per eventuali migliorie.

Gli step successivi prevedono l'uso di due tipologie di metriche (85):

- Metriche transazionali
  I KPI maggiormente adoperati tra tutti quelli che appartengono a questo gruppo sono verosimilmente il CSAT (Customer Satisfaction) e il CES (Customer Effort Score), che quantificano una determinata interazione e restituiscono l'apprezzamento dei clienti in un particolare momento. La limitazione di queste metriche è dovuta al fatto che non viene detto nulla sull'interazione complessiva o sulla sequenza delle diverse interazioni.

- Metriche di relazione
  Il KPI più inerente a questo gruppo di metriche è sicuramente l'NPS (Net Promoter Score), indicatore che esprime la propensione di un cliente a consigliare un prodotto, un servizio o un brand a parenti e conoscenti. Da un punto di vista analitico, esso coincide dunque con la probabilità di raccomandazione di un determinato marchio.

Un altro indicatore piuttosto utilizzato è il tasso di fidelizzazione del cliente (o customer retention), il quale esprime il numero di clienti che ritornano a comprare dopo aver acquistato già in passato, la qualità percepita del servizio clienti e le prestazioni del prodotto, perciò in sostanza viene misurato il grado di fedeltà del cliente.

In sintesi, le metriche di relazione fungono un po' da complementare delle metriche transazionali ovvero con esse è possibile osservare dall'inizio alla fine la storicità della relazione tra cliente e brand, quindi la completa interazione fra le due parti.

4.2.2 Requisiti per garantire l'efficacia dei segmenti

Non basta solamente definire e identificare i segmenti ma è necessario assicurarne anche l'efficacia, perciò devono essere soddisfatti diversi requisiti tra cui (60):

- Misurabilità
  Bisogna determinare quanto un particolare segmento spenderà per il relativo prodotto/servizio di un'azienda (esso potrebbe per esempio essere composto da clienti più propensi ad acquistare in periodi di saldi e promozioni e chi invece no), dunque è necessario stabilire il suo potenziale valore;

- Accessibilità
  Oltre a comprendere i clienti bisogna anche saperli raggiungere. Per far questo, molto spesso è utile comprendere le abitudini e gli stili di approccio che li portano ad acquistare un certo prodotto;

- Concretezza
  Una volta individuato il segmento, bisogna che le persone al suo interno non siano solamente interessate all'offerta proposta ma che siano anche in grado di acquistare il relativo prodotto.

  Ad esempio, si supponga che un retailer venda oggetti di lusso e che incentivi i clienti al loro acquisto. Probabilmente, molte persone saranno particolarmente interessate alla merce ma ciò non significa che tutti abbiano le risorse economiche per potersele permettere;

- Unicità
  Ogni segmento dev'essere unico e diverso dagli altri. Si supponga che la relativa targetizzazione abbia evidenziato abitudini d'acquisto simili tra due gruppi di persone. Conviene definire un segmento solo piuttosto che definire due segmenti separati.

4.3 Personalizzazione dei messaggi di comunicazione (volantino personalizzato)

In base alla fase della relazione tra il cliente e la GDO (nuovo cliente, cliente disaffezionato a rischio abbandono, cliente che riduce la sua varietà di spesa, cliente premium, etc), il messaggio da destinare a egli è diverso.

Un'idea per far ciò consiste nel comunicargli un'offerta che si adatti il più possibile alle sue esigenze. La comunicazione è sicuramente più efficace mediante la scrittura di un messaggio specifico/personalizzato piuttosto che uno generale.

La personalizzazione rappresenta l'obiettivo ultimo della segmentazione del pubblico di riferimento, e per capire in cosa consiste, bisogna inanzitutto che la relativa azienza o brand si ponga le seguenti domande prima di poter procedere con l'attività:

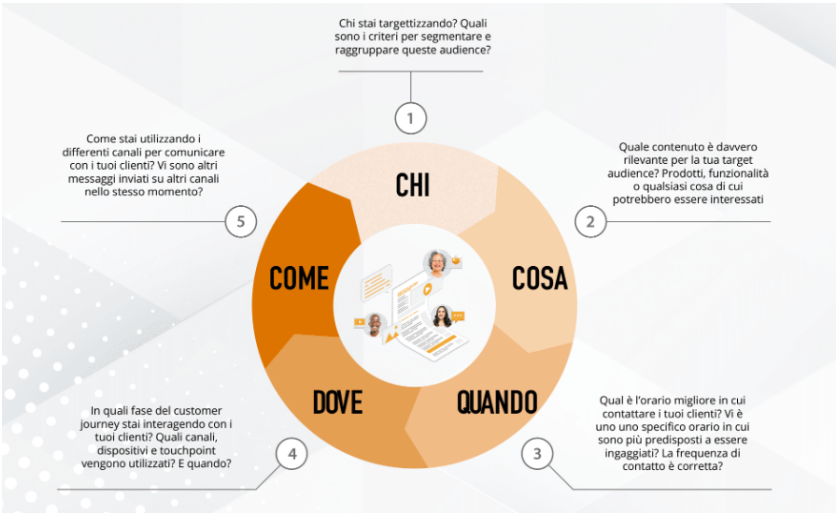


Figura 4.8: Puzzle personalizzazione
(68)

4.3.1 Aree trascurate nella personalizzazione delle comunicazioni

Al giorno d'oggi alcune aziende pensano esclusivamente a catturare un bacino di clienti che sia il più ampio possibile. In questo modo, non viene però fatto affidamento alle best practices utili per ottenere personalizzazioni dei messaggi che siano efficaci e ottimali. Infatti, non è raro che vengano trascurati alcuni degli aspetti seguenti (75):

- Orari e canali
  Aprire le email è un'operazione che viene eseguita in ogni istante della giornata e dovunque. Tuttavia, ci sono momenti in cui un cliente è particolarmente impegnato.

  Un'idea potrebbe consistere nell'inviare email nelle ore di punta oppure sfruttare (attraverso specifiche soluzioni) i dati sui propri clienti per venire a conoscenza del momento esatto in cui essi aprono usualmente le email.

  Per evitare momenti della giornata inopportuni, il messaggio potrebbe dunque essere recapitato in un range temporale consono, ma normalmente una volta identificata la target list avviene un unico invio a tutti i destinatari.

  Ciò accade poichè l'informazione di "quando inviare il messaggio" dovrebbe essere trasmessa al software di marketing automation e da quest'ultimo gestita, cosa abbastanza complicata;

- Messaggi poco rilevanti a causa di similarità
  Questo problema si manifesta con una certa frequenza soprattutto nei casi in cui un'azienda lavora con molteplici piattaforme. Ciò è dovuto al fatto che non sempre comunicare tanto significa creare alto interesse nel cliente, anzi il rischio di sovrapposizione di messaggi tra campagne differenti è particolarmente elevato.

  Elemento che risulta fondamentale per risolvere tale situazione è la CDP (Customer Data Platform). Essa permette di unificare tutti i dati dei clienti e in seconda battuta di contattare quest'ultimi nel momento opportuno attraverso una coordinamento delle varie comunicazioni che sono state programmate.

In questo modo, il rischio di sovrapposizione cala enormemente e la personalizzazione avrà più chance di successo;

- Segmentazioni dispersive dei clienti
  Un errore assolutamente da evitare per le aziende è quello di generare segmentazioni enormi e dispersive che cercano di catturare la più alta percentuale di persone.

  Anzi, impiegando specifiche metriche e insight, l'obiettivo è raggiungere risultati migliori targetizzando i clienti soprattutto secondo i loro interessi e i loro comportamenti, e questo lo si fa adoperando segmenti di pubblico più accurati e dettagliati;

- ANALISI RFM
  Nel caso in cui ci sia un cliente a rischio abbandono che potrebbe necessitare di un ulteriore incentivo per tornare ad interagire con il proprio marchio, un'idea per non perderlo consiste nell'offrirgli sconti, promozioni e perchè no anche iniziative di cross-selling e upselling.

  Ciò lo si fa attraverso un modello di marketing che può essere impiegato per la segmentazione avanzata automatizzata ovvero l'analisi RFM (Recency Frequency Monetary Value). In questo modo, tutti saranno maggiormente soddisfatti di ricevere contenuti più personalizzati compresi i clienti fidelizzati.

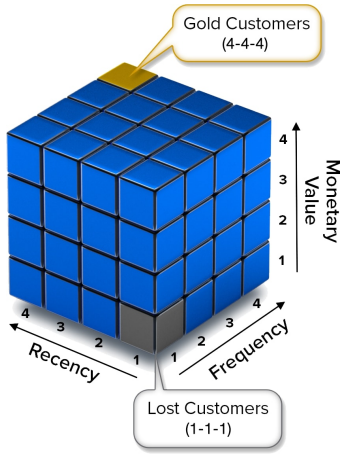


Figura 4.9: RFM
(34)

Nell'analisi RFM, le soglie possono essere definite o in modo fisso (si decidono dei valori immutabili che facciano da discriminante nella classificazione dei clienti) oppure in modo dinamico. Per attuare quest'ultimo approccio, generalmente si fa uso dei quantili quindi non vengono più fissate le soglie ma le numerosità.

I quantili, detti indici di posizione di non centralità, suddividono il set dei dati (che devono essere ordinati) in un numero specifico di parti uguali. Quest'ultimo viene solitamente settato a 100, 10, 4 o 2 e quindi si parlerà rispettivamente di percentile, decile, quartile o mediana.

Prendendo come riferimento i quartili, il primo quartile Q1 è un valore tale per cui il 25% dei dati ordinati è minore o uguale a Q1 (per questo motivo, viene anche chiamato 25-esimo percentile e indicato con $P_{0,25}$). Il secondo quartile Q2 (50-esimo percentile) corrisponde alla mediana. Il terzo quartile Q3 è un valore tale per cui il 75% dei dati ordinati è minore o uguale a Q3 (viene anche chiamato 75-esimo percentile e indicato con $P_{0,75}$).

Per calcolare i quartili, si applicano i passaggi seguenti (generalizzabili anche per gli altri indici di posizione) (51):

1. Si applica l'ordinamento crescente degli n dati;
2. Si calcola il prodotto k = np (per il primo quartile p = 0,25, per il secondo quartile p = 0,5 e per il terzo quartile p = 0,75);
3. Se k è un intero, il quartile si ottiene calcolando la media del k-esimo e del (k+1)-esimo valore dei dati ordinati;
4. Se k non è intero, si arrotonda k per eccesso al primo intero successivo e si sceglie come quartile il corrispondente valore dei dati ordinati.

Oltre ai quantili, si fa uso anche dei concetti di distribuzione e di statistica d'ordine. Si supponga di voler generare n replicazioni indipendenti di un esperimento per ottenere un campione casuale di dimensione n dalla distribuzione di una variabile casuale X, con funzione di ripartizione F e funzione di densità f: $(X_1, ..., X_n)$.

Sia $X_{(k)}$ il valore k-esimo più piccolo di $(X_1, ..., X_n)$. $X_{(k)}$ viene chiamata k-esima statistica d'ordine. Le statistiche d'ordine estremo sono i valori minimo

e massimo (41):

$$X_{(1)} = min\{X_1, ..., X_n\}$$
$$X_{(n)} = max\{X_1, ..., X_n\} \tag{4.1}$$

4.3.2 Engine di raccomandazione e comportamento d'acquisto dei clienti

Inoltre, per personalizzare un messaggio si potrebbe lavorare su altri aspetti che coinvolgono prevalentemente le preferenze d'acquisto. Tali preferenze d'acquisto potrebbero essere determinate da un algoritmo specifico (ad esempio l'engine di raccomandazione), che può dare suggerimenti per la creazione di messaggi mirati.

C'è una parte completamente automatica (Reco Engine) che genera un output puramente data driven, quindi esso prende in pasto dei dati e fornisce delle raccomandazioni perlopiù individuali. Tuttavia, ci sono delle regole di business o di senso pratico che potrebbero modificare in qualche modo il risultato dell'algoritmo. Per esempio, si consideri la situazione in cui si propone della carne ad un vegetariano oppure degli alcolici ad un minorenne.

Inoltre, ci sono logiche di tipo promozionale per cui magari la regola invoglia a comprare un prodotto che non ha un margine di profitto elevato per l'azienda, perciò al cliente verrà consigliato un prodotto simile ma con marginalità superiore. Poi c'è la parte di delivery (distribuzione delle raccomandazioni) che può avvenire via email, via app e quant'altro. Infine, c'è la valutazione e il monitoraggio dell'algoritmo.

È possibile comporre la proposta di acquisto ad ogni cliente fidelizzato combinando offerte basate su:

1. Prodotti da lui acquistati frequentemente assieme (market basket analysis);

2. Ripetitività dell'acquisto;

3. Similitudini di comportamento (prodotti "consigliati per te").

Tutti i punti hanno lo stesso vantaggio ovvero l'ottimizzazione automatica del relativo modello impiegato ma presentano le rispettive limitazioni:

1. L'acquisto in negozio ha dinamiche differenti rispetto a quelle online perciò c'è il rischio di catturare associazioni ovvie (per questo ed altri motivi la market basket analysis non è molto efficace nella realtà);

2. I filtri potrebbero non funzionare vista la ripetitività degli articoli proposti;

3. L'attivazione richiede la presenza di eventi solo online (dettagli, aggiunta al carrello, etc).

4.3.3 Collaborative filtering

Il concetto più comune e ricorrente che ci si trova spesso a trattare quando si parla di personalizzazione dei messaggi è il collaborative filtering. Esso presenta due modalità che sono illustrate qui di seguito (solitamente viene usato un mix delle due):

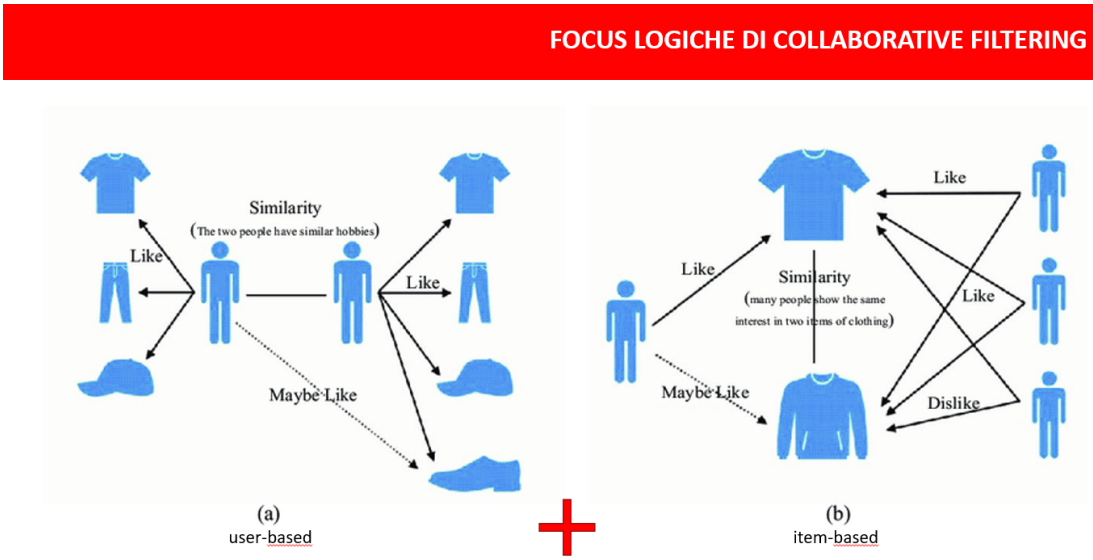


Figura 4.10: Collaborative filtering
(68)

- User-based: per riprendere l'esempio della figura, si consideri un utente che compra t-shirt, pantaloni e berretto ed un altro cliente che compra le stesse cose ma in più anche le scarpe. Sulla base della somiglianza che queste due persone hanno, viene suggerito l'acquisto delle scarpe al primo.

- Item-based: si considerino tre persone; alle prime due piacciono sia t-shirt che felpa mentre alla terza non piace nessuna delle due. In base a quanto è concorde il giudizio delle persone, viene in questo caso proposta anche la felpa ad una quarta persona che è oggetto dell'indagine, in quanto l'algoritmo sottostante suppone che ci sia qualche associazione tra la t-shirt e la felpa stessa.

Tra gli strumenti maggiormente impiegati nel collaborative filtering vi è la matrice user-rating, in cui vengono espresse le preferenze di molteplici clienti rispetto a differenti prodotti (item). Essa è una matrice m×n, in cui le m righe indicano i diversi clienti mentre le n colonne identificano i prodotti. La cella $r_{u,i}$ corrisponderà quindi alla valutazione attribuita al prodotto i da parte del cliente u. (64)

Per ottenere una previsione precisa nell'approccio user-based, è necessaria l'introduzione di una misura di similarità userSim(u,n), con la quale è possibile pesare i rating dei clienti più simili ad u. Un modo ricorrente per ottenere quest'ultima consiste nell'utilizzare il coefficiente di correlazione di Pearson, che confronta i rating per tutti i prodotti valutati sia dal cliente target sia dal "vicino" n:

$$userSim(u,n) = \frac{\Sigma_{i \subset CR_{u,n}}(r_{ui} - \bar{r_u})(r_{ni} - \bar{r_n})}{\sqrt{\Sigma_{i \subset CR_{u,n}}(r_{ui} - \bar{r_u})^2}\sqrt{\Sigma_{i \subset CR_{u,n}}(r_{ni} - \bar{r_n})^2}}, \tag{4.2}$$

dove r_u e r_n sono rispettivamente i rating del cliente target u e del cliente con preferenze affini n, $\bar{r_u}$ e $\bar{r_n}$ rappresentano rispettivamente la media delle valutazioni manifestate dal cliente u e il cliente n mentre CR fa riferimento ad un set di prodotti (item) valutati sia da u che da n. Il coefficiente di Pearson assume sempre valori compresi tra -1 e 1, dove 1 indica completa similarità tra i due clienti mentre -1 indica completa dissimilarità.

La previsione del rating per il prodotto i sarà dunque:

$$pred(u,i) = \bar{r_u} + \frac{\Sigma_{n \subset neighbors} userSim(u,n) \times (r_{ni} - \bar{r_n})}{\Sigma_{n \subset neighbors(u)} userSim(u,n)}. \tag{4.3}$$

Invece, per quanto riguarda l'approccio item-based, itemSim(i,j) indica la similarità tra i prodotti (item), che può essere ottenuta attraverso l'adjusted-cosine

similarity. Tale metrica, oltre ad essere la più usata, è anche la più precisa e viene ricavata utilizzando tutti i clienti che hanno espresso una valutazione sia per i che per j:

$$itemSim(i,j) = \frac{\Sigma_{u \subset RB_{I,J}}(r_{ui} - \bar{r_u})(r_{uj} - \bar{r_u})}{\sqrt{\Sigma_{u \subset RB_{i,j}}(r_{ui} - \bar{r_u})^2}\sqrt{\Sigma_{u \subset RB_{i,j}}(r_{uj} - \bar{r_u})^2}}, \tag{4.4}$$

dove RB è il set di clienti che hanno espresso una valutazione sia per i prodotti i che per j. Analogamente a quanto accade per il coefficiente di Pearson, anche questa metrica assume valori tra -1 e 1.

La previsione per il cliente u e il prodotto i si ottiene dalla somma pesata dei rating di u per i prodotti più simili ad i:

$$pred(u,i) = \frac{\Sigma_{j \in ratedItems(u)} itemSim(i,j) \times r_{ui}}{\Sigma_{j \in ratedItems(u)} itemSim(i,j)}. \tag{4.5}$$

A differenza delle formule 4.2 e 4.3, la correzione per la media è assente. Ciò è dovuto al fatto che i rating giungono dal cliente. Inoltre, c'è chiarezza sul fatto che gli algoritmi item-based sono più precisi rispetto a quelli user-based.

L'algoritmo che sta alla base della personalizzazione dei messaggi di comunicazione è affidabile ed efficiente perchè non dipende dal contesto. Idealmente funziona in scenari B2C (Business to Consumer) quindi tutto ciò che è meccanismo di retail ma funziona anche in scenari B2B (Business to Business) come ad esempio la vendita all'ingrosso.

Ovviamente tutto questo ha dei costi, i quali sono da affrontare durante le fasi seguenti:

- Preparazione dei dati;
- Training/retraining del modello;
- Acquisizione delle raccomandazioni.

4.3.4 Diritto alla privacy

Uno dei temi più discussi ad oggi è quello che riguarda il diritto alla privacy. Si pensi alle classiche spunte che si mettono sui permessi, che nell'ordine sono rispettivamente consenso alla profilazione, marketing diretto e comunicazione dei

dati a soggetti terzi. Molto importanti (soprattutto la prima) perchè se un cliente non vuole essere profilato tutto ciò comporta a cascata una serie di conseguenze.

A seconda del contesto in cui ci si trova, se un cliente ha negato la sua profilazione può non finire neanche in UCV (e quindi sparisce fin da subito) oppure finisce in UCV ma in seguito verrà filtrato prima del processo di marketing automation.

Questo dipende molto dal DPO (Data Protection Officer) ovvero la figura interna/esterna che si occupa di garantire le regole di protezione dei dati. Tutti i non profilabili diventano un'entità anonima che viene aggregata, dunque anche i relativi indicatori verranno calcolati in forma aggregata.

Questo tema di privacy è perciò molto rilevante e il giusto equilibrio può essere raggiunto solamente tramite una gestione attenta e rispettosa dei dati del cliente, affinchè non si sperperi il credito di fiducia guadagnato.

In ogni caso non è possibile conservare informazioni di profilazione relative ad un cliente (anche se lui ha dato i vari consensi) per una durata superiore ad un limite prefissato, perciò con la tabella si procede a riempimento finchè non si raggiunge la dimensione massima, utilizzando la tecnica della coda circolare.

Un altro discorso di rilievo è legato alla cifratura dei dati, i quali arrivano in chiaro ma nel momento in cui entrano all'interno della CDP (Customer Data Platform) devono essere cifrati.

Un esempio (che è già stato fatto nel paragrafo "KPI standard della UCV") è la data di nascita, la quale è considerata un attributo sensibile poichè da essa si potrebbe ricostruire il codice fiscale.

Questa cifratura deve essere inoltre invertibile ovvero nel momento in cui le varie informazioni vengono usate per contattare i clienti, esse devono tornare in chiaro.

4.4 Analisi rischio abbandono cliente

4.4.1 Churn Rate: definizione e cause

In ambito Retail, il Customer Churn Rate mostra la percentuale di clienti che smette di frequentare una particolare GDO (supermercato) nell'arco di un determinato periodo temporale, che può essere il mese, il trimestre, il quadrimestre, il semestre

o l'intero anno (quest'ultimo descrive comportamenti consolidati del cliente ed è in grado di annullare gli effetti di stagionalità nelle abitudini d'acquisto rilevate).

Il Churn Rate si calcola in questa maniera:

$$(\textit{Clienti persi in un periodo} \;/\; \textit{Clienti totali a inizio periodo}) * 100 \qquad (4.6)$$

Come si evince dalla formula, è importante specificare su quale periodo temporale effettuare l'analisi. La scelta varia a seconda dell'ambito di business e dal settore merceologico di riferimento in cui ci si trova, ma alla fine la sostanza è circa la stessa.

L'abbandono di clienti è un problema che si manifesta in tutti i settori e non è così difficile che il Churn Rate sia molto alto anche per le aziende leader nel loro settore.

Tra le cause di abbandono più frequenti si possono individuare (54):

- Prezzo: nel momento in cui due prodotti a confronto danno le stesse soddisfazioni, è normale che il cliente scelga con maggior probabilità l'opzione più economica sul mercato;
- Bassa qualità: un prodotto/servizio che manifesta imperfezioni, irregolarità o anomalie stimola la caccia ad altre soluzioni;
- Scarsa customer experience: con questo termine si vuole indicare il supporto clienti pre e post vendita e più nello specifico la qualità e velocità del servizio offerto, la fluidità dell'esperienza e la quantità di prodotti e relativi dettagli per assicurare acquisti maggiormente consapevoli;
- Cause non prevedibili: trasloco del cliente o apertura di nuovi competitor più vicini/economici che offrono le stesse soluzioni.

Per l'ultimo motivo, non ha del tutto senso usare questo tasso di rischio anche se il mercato è piuttosto interessato alla sua analisi. Il motivo che spinge a non trascurare del tutto il churn rate risiede nel fatto che le aziende possono rischiare di rallentare la propria crescita a livello di mercato o addirittura nei casi più estremi rischiare di chiudere e fallire.

L'analisi del Churn Rate permette non solo di frenare l'abbandono dei clienti, ma anche di valutare lo stato di salute del business attraverso l'utilizzo di indicatori chiave.

4.4.2 Approccio da seguire per ridurre il Churn Rate

Affinchè si possa ridurre il Churn Rate è essenziale riconoscere i fattori che scaturiscono l'origine di un suo alto valore, dopodichè si cerca di agire contrastando e limitando questo trend (54) - (10):

- Investire sulla qualità del prodotto/servizio: la concorrenza potrebbe dimostrarsi particolarmente ostica da un punto di vista tecnologico e ciò caratterizza uno degli aspetti più importanti per soddisfare e mantenere i clienti;
- Coinvolgere i clienti: sulla base di determinate necessità, bisogna dare supporto ai clienti mediante personalizzazioni di offerte e investimenti opportuni sui reparti marketing, vendite e customer care;
- Customer Satisfaction: richiedere in modo diretto ai clienti alcuni feedback e commenti sulla loro esperienza, in modo tale da individuare e valutare gli aspetti chiave della loro soddisfazione o insoddisfazione;
- Analisi predittiva e customer retention proattiva: si impiegano tecniche di AI per prevedere ed affrontare le cause che possono decretare l'abbandono della clientela.

4.4.3 Definizione di cliente perso e suo potenziale recupero

Se fosse necessaria la registrazione del cliente a un servizio, allora il Churn Rate sarebbe spesso impiegato nel settore di competenza. Per l'e-commerce ad esempio non si può dire lo stesso. Infatti, nei negozi online non è facile comprendere se un cliente è perso a tutti gli effetti oppure no.

Questo perchè esso non richiede quasi mai un'esplicita disiscrizione al sito perciò potrebbe sempre tornare ad acquistare. Tale situazione porta alla creazione di enormi database costituiti in gran parte da persone non più interessate da tempo ma troppo pigre per disiscriversi. (83)

In alcune situazioni è fisiologico perdere un cliente: se in un periodo della sua vita ha avuto sempre necessità di particolari prodotti, non vuol dire che esso debba continuare a procurarseli per sempre.

Inoltre, per alcuni prodotti i cicli di vendita sono più lunghi che per altri (per esempio, in categorie quali gli occhiali da vista o i mobili per la casa è consuetudine che i clienti lascino passare diverso tempo tra due acquisti successivi) quindi non si può affermare prematuramente di essere di fronte ad un abbandono certo.

Tuttavia, in altri casi i clienti potrebbero smettere di acquistare anche se la loro intenzione è quella di continuare a farlo. Ciò potrebbe essere dovuto alla forte concorrenza oppure alla presenza di nuovi punti vendita nei pressi delle loro residenze.

Individuare quali persone sono ancora potenzialmente interessate all'acquisto permetterà di organizzare un efficiente piano marketing di recupero. Un'azione opportuna consiste nella segmentazione addizionale della lista di possibili clienti persi, affinchè si possa individuare il potenziale che ci sta dentro.

Questo lo si può fare attraverso una segmentazione RFM (Recency Frequency Monetary Value). Per prima cosa si individua il segmento con alta recenza (ovvero tutti quei clienti il cui ultimo acquisto è stato effettuato molto tempo addietro) in modo da selezionare i clienti persi.

Dopodiché bisogna filtrare ulteriormente la lista eliminando coloro che hanno fatto meno di x acquisti e che hanno speso poco. Questo consente di avere un set di clienti che in passato sono già stati fidelizzati ma su cui ora bisogna investire per un loro ritorno. (83)

A prescindere dalla strategia che si vuole utilizzare, l'importante è riconoscere i clienti che hanno manifestato nel tempo l'approvazione dei diversi prodotti. Queste relazioni di fiducia con i clienti sono fondamentali e vanno coltivate in modo continuo e duraturo.

4.4.4 Meglio mantenere piuttosto che acquisire i clienti

Una percentuale molto alta di Customer Churn può creare parecchi costi inaspettati per l'azienda. Perdere clienti porta sempre a una perdita di entrate, ma questo non è l'unico aspetto su cui concentrarsi.

Infatti, molte aziende tendono a incrementare i propri investimenti nell'acquisizione di nuovo pubblico piuttosto che cercare di recuperare quei clienti che un tempo erano abituali ma che ora sono a rischio abbandono.

I costi del primo investimento sono decisamente maggiori rispetto al costo di mantenimento di una base clienti già esistente e oltre a ciò un cliente di lunga data vale molto di più di un cliente appena acquisito.

Tuttavia, per ottimizzare ulteriormente la Customer Experience è fondamentale coinvolgere il servizio clienti anche in seguito all'acquisto, ovvero bisogna aver la capacità di fornire risposte adeguate e veloci nel momento in cui nascono delle complicanze.

Nel presente, le possibilità sono molteplici: ci si può rivolgere fisicamente ad un addetto alle vendite in negozio, si può richiedere assistenza online tramite chat e profili di social media, e così via.

In ogni caso, nel momento in cui i clienti condividono le loro opinioni e domande su tutti i canali dell'azienda, l'importante è essere immediati nel fornire loro una risposta adeguata. (10)

4.4.5 Churn Analysis: prevenire invece che curare

La Churn Analysis è un'analisi di previsione che mediante specifici modelli permette di rintracciare in modo automatizzato i clienti che presentano una probabilità elevata di abbandono, con lo scopo di agire in anticipo (seguendo le azioni più indicate per ogni livello di probabilità) per impedirne la migrazione presso un diverso competitor.

In ambito Retail, grazie all'analisi delle transazioni e delle informazioni disponibili sulla clientela (elaborate con l'impiego di sistemi di CRM), l'azienda può monitorare il grado di soddisfazione del cliente e agire di conseguenza per incrementarlo, con l'obiettivo di scongiurare il suo abbandono. (11)

4.4.6 Algoritmi di classificazione binaria

Per calcolare il Churn Rate si utilizzano prevalentemente algoritmi di classificazione binaria. Quest'ultimi consistono in metodi di apprendimento supervisionato in cui

una risposta categorica (la variabile Y) assume 2 possibili valori che sono 0 o 1 (un cliente non abbandona oppure abbandona).

Essa deve essere prevista a partire da un vettore X di variabili esplicative (frequenza, valore di spesa, etc), impiegando una funzione di predizione g. In tal senso, g classifica l'input X in una delle due classi possibili. Per questo motivo, g viene chiamata funzione di classificazione binaria o semplicemente classificatore binario.

Come con qualsiasi tecnica di apprendimento supervisionato, l'obiettivo è ridurre al minimo la perdita o il rischio atteso

$$l(g) = E[Loss(Y, g(X))] \tag{4.7}$$

per qualche funzione di Loss (Loss(y;$\hat{y}$)), che quantifica l'impatto della classificazione di una risposta y tramite $\hat{y}$ = g(x).

Per misurare la performance di un classificatore su un set di training o di test, è conveniente introdurre la nozione di Confusion Matrix, indicata con M, dove il (j;k)-esimo elemento di M conta il numero di volte in cui, per i dati di addestramento o test, la classe effettiva (osservata) è j mentre la classe predetta è k.

Per ogni classe j, a volte è utile dividere gli elementi di una confusion matrix in 4 gruppi:

1. True Positive: $tp_j = M_{jj}$;

2. False Positive: $fp_j = \sum_{k \neq j} M_{kj}$ (somma per colonna);

3. False Negative: $fn_j = \sum_{k \neq j} M_{jk}$ (somma per riga);

4. True Negative: tn_j = n - fn_j - fp_j - tp_j.

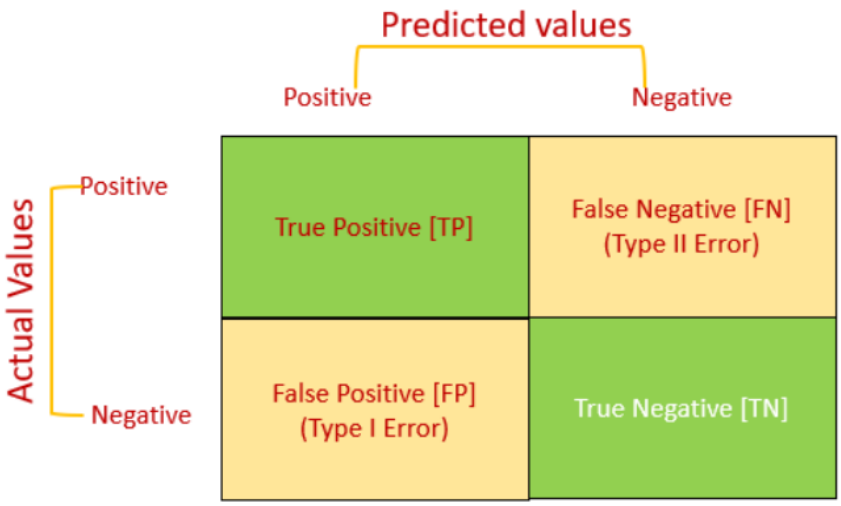


Figura 4.11: Confusion Matrix
(25)

Per decidere se l'osservazione debba essere classificata come 0 o 1, si interpreta l'output dell'algoritmo scegliendo una soglia di classificazione e si mettono a confronto queste due misure. Ogni osservazione con punteggio superiore alla soglia viene quindi assegnata alla classe 1 mentre i punteggi inferiori alla soglia vengono assegnati alla classe 0.

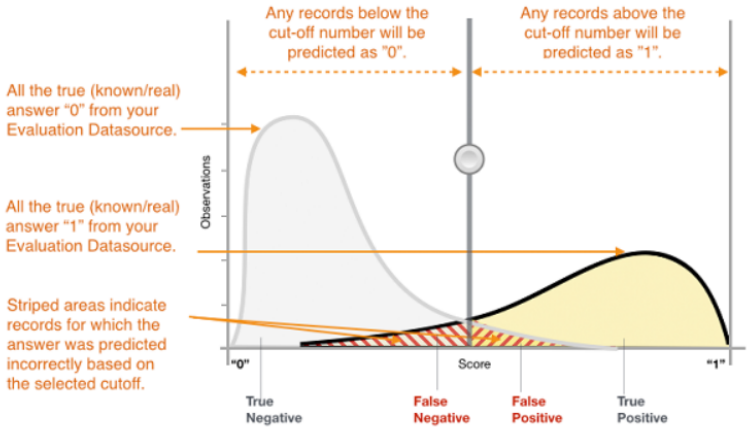


Figura 4.12: Classificazione binaria
(24)

In aggiunta, con gli elementi della confusion matrix è possibile definire diverse metriche:

- $Errore_j = \frac{fp_j+fn_j}{n}$;
- $Accuratezza_j = 1 - errore_j = \frac{tp_j+tn_j}{n}$ (la frazione di oggetti correttamente classificati);
- $Precisione_j = \frac{tp_j}{tp_j+fp_j}$ (la frazione di tutti gli oggetti classificati come j che effettivamente sono oggetti j);
- $Recall_j (o\ Sensitivit\text{à}) = \frac{tp_j}{tp_j+fn_j}$ (la frazione di tutti gli oggetti j che vengono correttamente classificati come tali);
- $Specificit\text{à}_j = \frac{tn_j}{fp_j+tn_j}$ (la frazione di tutti gli oggetti non j che vengono correttamente classificati come tali);
- $F_{\beta,j} (F_\beta\ score) = \frac{(\beta^2+1)tp_j}{(\beta^2+1)tp_j+\beta^2 fn_j+fp_j}$ (una combinazione di precisione e recall).

Fatta quest'ampia premessa, è ora possibile definire gli algoritmi di classificazione binaria più rilevanti (65):

- Logistic Regression: il modello di regressione logistica (logit) è un modello lineare generalizzato in cui, condizionatamente a un vettore x di KPI p-dimensionale, la risposta casuale Y ha una distribuzione Ber(h($x^T > \beta$)) con h(u) = $1/(1+e^{-u})$.

  Il parametro β viene appreso dai dati di training massimizzando la verosimiglianza delle risposte di addestramento o, equivalentemente, minimizzando la versione supervisionata della cross-entropy training loss:

$$-\frac{1}{n}\sum_{i=1}^{n} ln\ g(y_i|\beta, x_i), \tag{4.8}$$

  dove $g(y = 1|\beta, x) = 1/(1 + e^{-x^T\beta})$ e $g(y = 0|\beta, x) = e^{-x^T\beta}/(1 + e^{-x^T\beta})$.

  In particolare, si ha che

$$ln\frac{g(y = 1|\beta, x)}{g(y = 0|\beta, x)} = x^T\beta. \tag{4.9}$$

In altre parole, il rapporto log-odds è una funzione lineare del vettore di KPI. Di conseguenza, il decision boundary $x : g(y = 0|\beta, x) = g(y = 1|\beta, x)$ è l'iperpiano $x^T\beta = 0$. Si noti che x in genere include la feature costante. Se quest'ultima è considerata separatamente, cioè $x = [1, \tilde{x}^T]^T$, allora il boundary è un iperpiano affine in $\tilde{x}$.

Supponiamo che l'addestramento su $\tau = \{(x_i, y_i)\}$ dia la stima $\hat{\beta}$ con il learner corrispondente $g_\tau(y = 1|x) = 1/(1 + e^{-x^T\hat{\beta}})$. Il learner può essere utilizzato come pre-classificatore da cui si ottiene il classificatore $\mathbf{1}\{g_\tau(y = 1|x) > 1/2\}$ o, equivalentemente,

$$\hat{y} := \underset{j\in[0,1]}{\operatorname{argmax}}\, g_\tau(y = j|x). \tag{4.10}$$

- Support Vector Machine (SVM): è un algoritmo di apprendimento supervisionato impiegato spesso in problemi di classificazione binaria, che ha lo scopo di individuare un iperpiano che separi in modo ottimale i clienti che non abbandonano da quelli che abbandonano.

L'iperpiano deve quindi avere il margine più ampio possibile tra le due classi, rappresentate solitamente con i segni + e -. Il termine margine indica la larghezza massima della linea parallela all'iperpiano che non possiede punti interni.

L'algoritmo può individuare un iperpiano del genere esclusivamente per i problemi separabili linearmente mentre per i problemi più pratici l'algoritmo massimizza il margine soft, il che permette di avere un numero basso di classificazioni sbagliate.

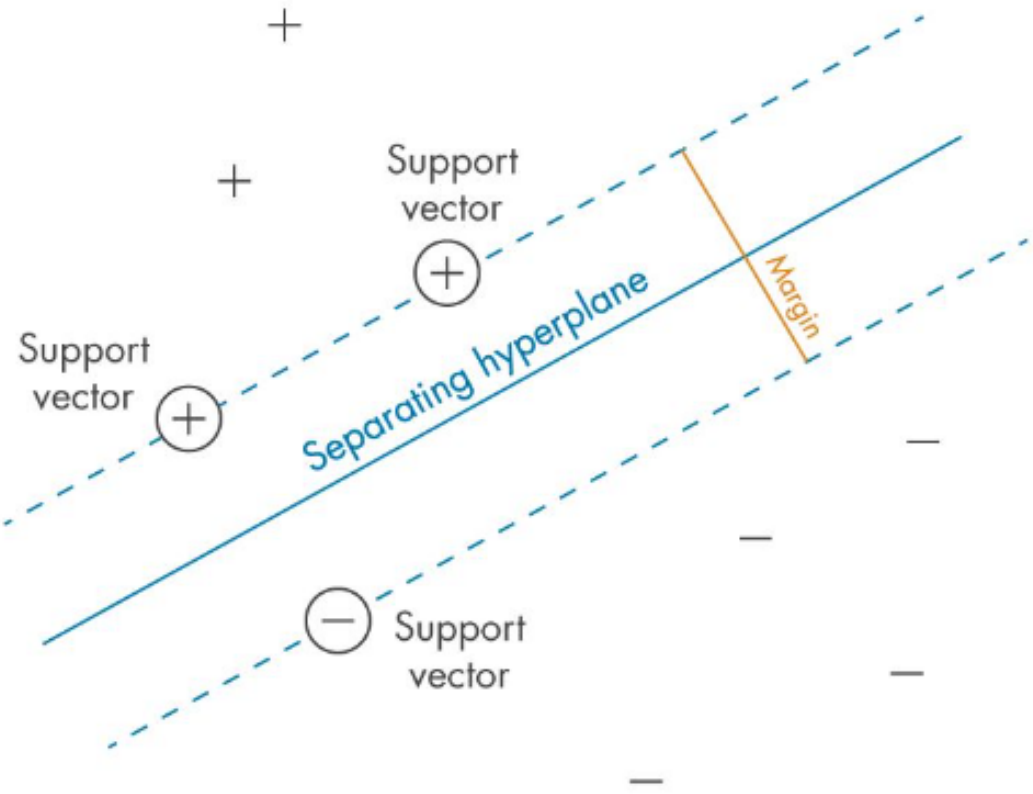


Figura 4.13: SVM
(35)

La posizione dell'iperpiano di separazione viene individuata da un sottoinsieme di osservazioni di training chiamato Support Vector. Da un punto di vista matematico, le SVM corrispondono ad una classe di algoritmi di ML definiti metodi kernel, in cui si è in grado di applicare una trasformazione ai KPI attraverso delle funzioni kernel. Quest'ultime mappano i dati su uno spazio differente e solitamente di dimensione maggiore.

Questo approccio mira a semplificare la separazione tra le classi in quanto i confini di decisione non lineari complessi vengono tradotti in confini lineari nello spazio di dimensione maggiore. A questo punto entra in gioco il "kernel trick", che consente di evitare trasformazioni esplicite dei dati e quindi un dispendio computazionale elevato.

Tipo di SVM	Kernel di Mercer	Descrizione
Radial Basis Function (RBF) o gaussiano	$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$	Addestramento di una classe. σ è la larghezza del kernel
Lineare	$K(x_1, x_2) = x_1^\mathsf{T} x_2$	Addestramento di due classi.
Polinomiale	$K(x_1, x_2) = \left(x_1^\mathsf{T} x_2 + 1\right)^\rho$	ρ è l'ordine del polinomio
Sigmoide	$K(x_1, x_2) = \tanh\left(\beta_0 x_1^\mathsf{T} x_2 + \beta_1\right)$	Si tratta di un kernel di Mercer solo per alcuni valori β_0 e β_1

Figura 4.14: Kernel SVM
(32)

- Decision Tree: durante la classificazione, un albero si concentra su una sequenza di scelte basate sul valore di uno o più KPI. Partendo dalla radice, le scelte attraversano l'intero albero fino alle foglie terminali.

  Dal punto di vista geometrico, esso (man mano che prosegue) suddivide lo spazio predittivo in più regioni, e dopo vari steps si avranno delle foglie con una maggior presenza di etichette relative alla classe 0 (il cliente non abbandona) piuttosto che alla classe 1 (il cliente abbandona). In questo modo, l'albero associa una determinata classe al vettore x di KPI del cliente in esame.

  Il vantaggio principale di un albero è la sua chiara interpretabilità umana. Inoltre, una parte molto importante è lo split dei dati, che avviene attraverso l'entropy e l'indice di Gini:

  – entropy:
  $$-\sum_{z=0}^{c-1} p_z log_2(p_z)$$
  dove p_z rappresenta la proporzione di dati che sono etichettati con la classe z. Viene calcolata la differenza di entropia tra i vari KPI prima dello split e dopo che quest'ultimo è stato eseguito, dopodichè viene scelto il KPI con il valore maggiore come nodo di scelta successivo nell'albero.

  – indice di Gini:
  $$\frac{1}{2}(1 - \sum_{z=0}^{c-1} p_z^2)$$
  dove p_z^2 rappresenta la probabilità che una tupla nel sottospazio precedente dell'albero appartenga alla classe z-esima.

Entrambe le misure di impurità hanno valore massimo quando le probabilità di classe sono uguali a 1/c (in questo caso c=2).

Quando nel sottospazio raggiunto si hanno soltanto oggetti di una determinata classe, questi indici raggiungono lo 0 quindi non ci sono diversificazioni di classe.

Invece, un valore positivo di entropia indica che non si è ancora arrivati ad un valore univoco di classe, perciò in quel caso bisogna valutare la classe con la più alta probabilità di associazione.

- Random Forest: per un vettore x di KPI si definisca $Z_b = g_{\mathcal{T}_b}(x), b = 1, ..., B$ iid valori di predizione, ottenuti da set di training indipendenti $\mathcal{T}_1, ..., \mathcal{T}_B$. Si supponga che $Var(Z_B) = \sigma^2$ per tutti i $b = 1, ..., B$. Allora la varianza del valore medio di predizione $\bar{Z_B}$ è uguale a σ^2/B.

  Se invece si usassero data sets bootstrapped $\{\mathcal{T}_b^*\}$, le variabili random corrispondenti sarebbero correlate Z_b. In particolare, $Z_b = g_{\mathcal{T}_b^*}(x)$ con $b = 1, ..., B$ sono identicamente distribuite ma non indipendenti, con qualche correlazione positiva a coppie ϱ. Vale che:

$$Var\bar{Z}_B = \varrho\ \sigma^2 + \sigma^2 \frac{1 - \varrho}{B}. \tag{4.11}$$

  Mentre il secondo termine della 4.11 tende a zero all'aumentare del numero di osservazioni B, il primo termine rimane costante.

  Questo problema è particolarmente rilevante per l'inserimento di alberi decisionali. Ad esempio, si consideri una situazione in cui un KPI fornisca un'ottima suddivisione dei dati.

  Tale KPI verrà selezionato e suddiviso per ogni $\{g_{\mathcal{T}_b^*}\}$ con b=1,...,B alla radice e di conseguenza ci si ritroverà con previsioni altamente correlate. In una situazione del genere, la previsione media non introdurrà il miglioramento desiderato nelle prestazioni del predittore bagged.

  Questa limitazione dei Decision Tree viene risolta dal metodo Random Forest, che cerca di mantenere una bassa distorsione e ridurre anche la varianza calcolando la media del risultato su una serie di alberi diversi.

La procedura alla base della Random Forest seleziona casualmente un sottoinsieme di KPI diversi e costruisce alberi diversi su campioni di addestramento bootstrap. Ogni albero verrà costruito considerando esclusivamente un sottoinsieme di KPI (cosa positiva, in quanto porta a un minore overfitting e impedisce ad un KPI forte di predominare su tutti gli altri).

Due parametri importanti del modello sono il numero di KPI per ogni albero e l'altezza dell'albero. La Random Forest migliora l'accuratezza rispetto alle previsioni fatte con un singolo albero decisionale, ma il suo output è molto più difficile da interpretare.

Quindi quando si deve studiare la classe di appartenenza di un nuovo cliente, si calcola la presunta classe con tutti gli alberi presenti nella random forest e successivamente viene considerata la classe che occorre maggiormente tra tutte le classi temporanee.

Si può anche studiare l'importanza dei KPI nel modello di classificazione, quindi vedere quali KPI sono più importanti per suddividere in porzioni lo spazio di dati, per poi definire una regione dove i dati presenti sono tutti o quasi della stessa classe e procedere quindi con la definizione della classe di appartenenza.

- Bayes' rule: classificare il vettore x di KPI relativo ad un cliente in base alla loro probabilità di classe condizionata è una cosa naturale da fare, specialmente in un contesto di apprendimento bayesiano. Nello specifico, la probabilità condizionata f(y|x) viene interpretata come una probabilità a posteriori della forma

$$f(y|x) \ \alpha \ f(x|y)f(y), \tag{4.12}$$

dove f(x|y) è la verosimiglianza di ottenere il vettore x di KPI dalla classe y mentre f(y) è la probabilità a priori di appartenere alla classe y. Facendo varie ipotesi su quest'ultima (ad esempio, tutte le classi sono a priori ugualmente probabili) e sulla verosimiglianza, si ottiene la a posteriori tramite la formula 4.12.

Una classe viene quindi assegnata a un vettore x di KPI secondo la probabilità a posteriori più alta, cioè la classificazione avviene secondo la regola decisionale

ottimale di Bayes:

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, f(y|x). \tag{4.13}$$

Inoltre, esiste una variante di questo metodo (chiamata Naive Bayes) che ha l'obiettivo di facilitare i calcoli, assumendo che l'effetto di un KPI su una data classe sia indipendente dai valori degli altri KPI. Tale assunzione è chiamata indipendenza condizionata delle classi.

Dunque, si supponga che un vettore $x = [x_1, ..., x_p]^T$ di p KPI debba essere classificato nella classe 0 o 1. Nel metodo Naive Bayes, la classe di funzioni approssimanti $\mathcal{G}$ è scelta in modo tale che $g(x|y) = g(x_1|y) \cdots g(x_p|y)$ ovvero tutti i KPI sono indipendenti (condizionatamente alla classe).

Assumendo una priori uniforme per y, la pdf a posteriori può essere scritta come segue:

$$g(y|x) \ \alpha \ \prod_{j=1}^{p} g(x_j|y) \tag{4.14}$$

dove le pdf marginali $g(x_j|y)$, j = 1,...,p appartengono ad una data classe di funzioni approssimanti $\mathcal{G}$. Per classificare x, semplicemente si considera la y che massimizza la pdf a posteriori non normalizzata.

4.5 Definizione del CLV (Customer Lifetime Value)

4.5.1 CLV: definizione e misurazione

Il Customer Lifetime Value (CLV) è una metrica che indica il profitto medio che ogni cliente genera durante tutto il rapporto commerciale con l'azienda. Ad un aumento del CLV di solito corrisponde anche un aumento della loyalty (fidelizzazione) al brand.

Prima di investire nell'acquisizione di nuovi clienti sarebbe meglio che ogni azienda dedicasse più risorse nel mantenere la base clienti già esistente. Con mantenere non si intende solamente condurre il cliente fino all'acquisto ma anche accompagnarlo in un secondo momento, evidenziando così una spiccata attenzione che può suscitare in lui una certa soddisfazione.

Per poter misurare con accuratezza e precisione il CLV, è importante individuare i touchpoint nei quali il cliente ha creato valore (con touchpoint si intende qualsiasi canale attraverso cui l'azienda e il cliente possono entrare in contatto) e in seguito calcolare i profitti per ciascuno di essi. Infine, bisogna sommare questi profitti per l'intera vita del cliente.

Identificare i touchpoint che generano più valore aiuta a capire dove mettere mano per ottimizzare la propria strategia e dove, invece, è necessario un cospicuo miglioramento.

Inoltre, il CLV cresce di importanza in proporzione alla longevità della relazione, ovvero rapporti più lunghi generano un valore più alto per l'azienda. Il tempo è perciò un fattore essenziale che deve essere sempre tenuto sotto osservazione.

Il CLV (durante un arco temporale specifico) si ottiene dalla seguente formula:

$$CLV = \textit{Ricavi totali generati dagli acquisti del cliente} - \textit{Costi di acquisizione del cliente} - \textit{Costi di retention (mantenimento)}. \tag{4.15}$$

Un aspetto su cui si concentra questa metrica è dunque il Cost-to-Serve ovvero il costo per fornire un servizio ad un cliente. Se quest'ultimo diventa troppo elevato, si rischia di ottenere minori profitti malgrado il CLV apparentemente alto del cliente in questione. Perciò è fondamentale trovare il giusto equilibrio. (52)

4.5.2 Approccio storico e approccio predittivo

Esistono due approcci per calcolare il CLV: approccio storico (4.15) e approccio predittivo. Nel primo caso si considerano gli acquisti e le transazioni effettuate dal cliente fino alla data odierna mentre nel secondo caso si può optare per una predizione degli acquisti futuri.

Per poter attuare un approccio predittivo è però necessario considerare dei fattori ulteriori come le transazioni mensili medie, l'ammontare medio di ogni transazione, il numero medio di mesi di fedeltà e il margine lordo medio.

Nonostante tutto, la scelta tra un approccio e l'altro dipende dall'organizzazione dell'azienda, dalle risorse disponibili e dal modello di business impiegato.

Sicuramente, il secondo approccio è più complesso perciò in diversi casi è meglio optare per l'approccio storico che è più semplice e può essere ad ogni modo una valida alternativa.

Per facilitare il calcolo del CLV possono essere definite e individuate delle macro-categorie per tipi di prodotti acquistati, per zona geografica o per compartecipazione di touchpoint.

Ad ogni modo, è possibile esprimere il CLV anche attraverso l'approccio predittivo:

$$CLV = \sum_{t=0}^{T} P(Active) \times \frac{(p-c)}{(1+d)^t} - AC. \tag{4.16}$$

In questo caso, il CLV corrisponde al valore attuale dei profitti futuri generati da un cliente per l'intera durata della sua relazione con la GDO.

P(Active) rappresenta la probabilità che il cliente sia attivo in futuro mentre p rappresenta i flussi di cassa per periodo (si fa riferimento alla media dei flussi passati generati da ogni cliente).

Per quanto riguarda c e AC, essi rappresentano i costi sostenuti per ogni cliente quindi sia i costi diretti (costi sostenuti per permettergli di fruire del prodotto/servizio offerto dalla GDO) sia i costi di acquisizione (che a livello individuale vengono ottenuti dividendo l'ammontare speso per una campagna di marketing per il numero di clienti acquisiti attraverso la campagna stessa).

Infine, $(1+d)^t$ rappresenta il tasso di attualizzazione, che è proporzionale al tasso di interesse pagato dalla banca sui conti correnti della GDO; alternativamente può essere usato il costo medio del capitale della GDO.

Discorso a parte deve essere fatto per P(Active): essa viene definita attraverso il modello di Cox, metodologia tipica dell'analisi di sopravvivenza applicata ad ogni specifico tempo t. Tale probabilità fa riferimento alla funzione di sopravvivenza e ad una serie di covariate.

$$h(t, X) = h_0(t)\ e^{\sum_{i=1}^{p} \beta_i X_i} \tag{4.17}$$

h(t,X) rappresenta il rischio al tempo t (sulla base delle covariate X), h_0 rappresenta la funzione di rischio di base, X rappresenta le variabili esplicative mentre β rappresenta il coefficiente assegnato alla variabile esplicativa. (47)

Per ogni cliente, P(Active) può essere trattata come una curva di sopravvivenza:

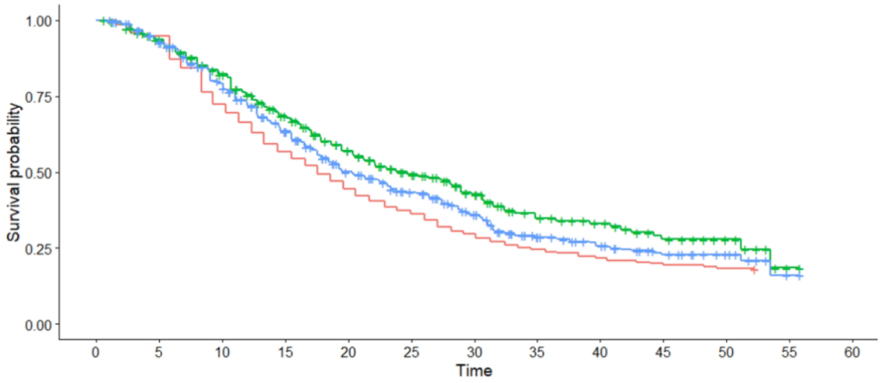


Figura 4.15: Curve di sopravvivenza
(26)

4.5.3 Strategia di ottimizzazione del CLV

Ci sono diverse best practices e strategie che consentono di migliorare e ottimizzare la metrica che definisce il Customer Lifetime Value. Qui di seguito vengono riportate quelle prioritarie ed essenziali (49):

- Comunicazione personalizzata
  Frequentemente, le aziende hanno database e CRM pieni di nomi, indirizzi mail e contatti che non descrivono praticamente nulla del cliente. Come ribadito più volte, è perciò fondamentale personalizzare la comunicazione con il cliente.

  Un beneficio può essere portato dal fatto di fare distinzione dei contatti a seconda del customer journey, dei touchpoint toccati e degli acquisti effettuati. In seguito a tali distinzioni, sarà possibile originare delle mail con offerte dedicate per ogni singolo cliente;

- Cross selling
  In base al genere di prodotto, cambia la relativa frequenza di vendita. Alcuni prodotti non hanno una vita molto lunga e sono dunque soggetti ad un costante ricambio mentre altre tipologie di prodotti vengono acquistate una volta per durare diverso tempo.

Per quest'ultima categoria, si cerca di sopperire alla bassa frequenza di vendita proponendo al cliente dei prodotti addizionali che siano coerenti con quello appena acquistato (concetto di cross selling);

- Sconti e coupon
  Scontistiche, promozioni e quant'altro sono sempre apprezzate dal cliente nel momento del suo prossimo acquisto, ma è necessario scegliere il momento adatto per proporre i coupons.

  Se lo scopo principale è quello di mantenere una base di clienti già esistente, normalmente si suggerisce di inoltrare la promozione non troppo presto (per non opprimere il cliente) ma neanche troppo tardi (per stimolare e conservare la voglia di una nuova compera). Dunque, è necessaro far trascorrere in media 15 giorni dall'ultimo acquisto;

- Email per stimolare i dormienti e per ringraziare dopo l'acquisto
  Attraverso un database aggiornato sui contatti, è possibile identificare in prima battuta i clienti "dormienti" e in un secondo momento stimolare loro con offerte allettanti che fungono da reminder per provare a riattivare esperienze passate.

  Ogni qualvolta si verifica un acquisto da parte dei clienti dormienti e non, è buona norma inviare una email di ringraziamento ai consumatori. E' una comunicazione che non richiede nessun tipo di sforzo, anzi esprime ed evidenzia la considerazione e la riconoscenza del brand nei confronti di ogni singolo cliente;

- Programmi fedeltà
  Attribuire punti e premi (sia offline sia passando per il digitale) rafforza il concetto di appartenenza, fedeltà e quindi relazione tra brand e cliente. Infatti, a quest'ultimo viene fornita la possibilità di ricevere sconti e privilegi che ad altri non sono concessi.

4.5.4 Importanza del CLV

Il calcolo del CLV applicato per diverse segmentazioni del cliente facilita soprattutto il processo decisionale aziendale. Conoscere questa metrica può informare l'azienda su diversi aspetti (53):

- La cifra spendibile per l'acquisizione di nuovi clienti ed avere nonostante tutto un rapporto proficuo;
- L'importo che può essere potenzialmente speso da un cliente medio nel tempo;
- Tipologie di prodotti che desiderano i clienti di alto valore;
- Prodotti che presentano la massima redditività;
- Relazioni con i clienti che determinano la maggior parte delle vendite;
- Tipologie di cliente più redditizie;
- Dettagli sul customer journey e Churn Rate.

4.6 Individuazione del TtNP (Time to Next Purchase)

Il TtNP o Time to Next Purchase è un indicatore impiegato nel mondo analisi dati (e soprattutto in ambito retail) per stimare il tempo medio che intercorre tra due acquisti successivi effettuati dal medesimo cliente.

Questo indicatore è importante sia per la comprensione dei comportamenti d'acquisto del cliente stesso ma anche per l'ottimizzazione delle strategie di marketing. A tutti gli effetti può essere definito come il complementare del Churn Rate.

Per ottenere l'indicatore che esprime il TtNP ci sono diversi passaggi da seguire (79) - (69):

- Trattamento dei dati
  Consiste nella pulizia, preparazione e trasformazione dei dati;
- Ingegneria delle caratteristiche
  Nell'effettuare questa operazione si impiega spesso la segmentazione RFM (Recency Frequency Monetary Value) per poter individuare in seguito cluster di clienti;

- Costruire modelli di machine learning
  Ci sono molteplici modelli di ML: Naive Bayes, LogisticRegression, RandomForestClassifier, GaussianNB, XGBClassifier, DecisionTreeClassifier, KNeighborsClassifier e così via.

  Per essere certi della stabilità del modello impiegato (su differenti set di dati) e per gestire eventuali rumori nel set di test selezionato si utilizza la cross validation. Essa procura il punteggio del modello utilizzando diversi test set. Il modello è stabile solo se la deviazione è bassa;

- Ottimizzazione degli iperparametri e selezione del modello
  Per migliorare il più possibile i modelli elencati al punto precedente, bisogna considerare l'ottimizzazione dei loro iperparametri. Mediante alcune funzionalità è possibile individuare i valori migliori per quest'ultimi, e in tal modo la scelta di un modello piuttosto che di un altro potrebbe cambiare.

  La selezione di un modello avviene attraverso la valutazione di diverse metriche che sono principalmente accuratezza, precisione, f1 score e recall. Generalmente si fa riferimento alle prime due.

4.6.1 Preparazione e selezione delle features

Per esprimere la stima sulla distanza in giorni al prossimo acquisto (ritorno del cliente), un qualsiasi algoritmo di machine learning richiede un set di features in input scelte ad hoc durante la fase di sviluppo e fine tuning del modello stesso, come ad esempio:

Nome Feature	**Significato**
Tot spesa	Importo in euro ultima spesa
Tot spesa x gg	Importo in euro negli ultimi x gg
Pdv	Punto vendita prevalente (solitamente per spesa/frequenza maggiore)
Giorni spesa week	Numero di gg dall'ultima spesa settimanale
Tot spesa week	Importo in euro dell'ultima settimana
Gg dist spesa osservata	Numero di giorni dall'ultima spesa
Frequenza x gg	Frequenza di acquisti calcolata negli ultimi x gg
Perc "nome reparto" x gg	Percentuale di prodotti proveniente da reparto "nome reparto" acquistati negli ultimi x gg
Intervallo medio riacquisto reale	Media degli intertempi d'acquisto reali

Tabella 4.1: Features del TtNP

Inoltre, può essere considerata anche una feature un po' più particolare come ad esempio l'indicatore che dà una misura dell'"accelerazione", ovvero la distribuzione nel tempo delle varie spese effettuate dal cliente. L'idea è considerare questi valori di spesa come dei punti da mettere su un piano cartesiano (l'asse x è il tempo mentre l'asse y è il valore di spesa del cliente) e si cerca di costruire un trend.

Il metodo più semplice consiste nel fare una retta di regressione (questo concetto vale anche per il Churn Rate). Il modello di regressione più elementare prevede una relazione lineare tra la risposta e un'unica variabile esplicativa. In particolare, si hanno le misure $(x_1, y_1), ..., (x_n, y_n)$ che giacciono approssimativamente su una linea retta.

Un modello semplice per questi dati prevede delle $\{x_i\}$ fisse e delle $\{Y_i\}$ random in modo tale che

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \;\; i = 1, ..., n, \tag{4.18}$$

per certi parametri sconosciuti β_0 e β_1 (assumendo gli $\{\epsilon_i\}$ indipendenti con valore atteso 0 e varianza σ^2). (65)

La linea di regressione (sconosciuta) è:

$$y = \beta_0 + \beta_1 \; x. \tag{4.19}$$

Le stime di β_0 e β_1 si ottengono minimizzando la funzione $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$:

$$\hat{\beta_0} = \bar{y} - \beta_1 \bar{x},\ \hat{\beta_1} = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\bar{y}}{\frac{\sum x_i^2}{n} - \bar{x}^2}. \tag{4.20}$$

La stima $\hat{\beta_1}$ mostra in questo caso specifico quanto il cliente sta accelerando con le sue spese.

Naturalmente, il modello di regressione semplice è piuttosto banale e non così accurato per i problemi più complessi. Ad esempio, si possono considerare i modelli di regressione non lineari (che permettono di individuare relazioni non lineari tra le variabili X e Y) o altri modelli ancor più sofisticati.

Tornando invece all'algoritmo del TtNP, le features da fornire in input richiedono l'aggregazione di diversi tipi di dato su finestre temporali che possono variare a seconda delle esigenze, ma si supponga abbiano durata massima di 365 + 30 giorni. La finestra di 30 giorni oltre all'anno di storico permette di lanciare la procedura di recovery del dato qualora si riscontri un'eventuale mancanza di scontrini.

Affiancando i dati di anagrafica, scontrinato e scontistica relativi agli ultimi 395 giorni di storico ed effettuando le opportune aggregazioni, viene creata una tabella che riporta per ogni giorno l'ultima spesa effettuata da ciascuno dei clienti (identificati tramite il proprio id conto). Inoltre, viene riportata una serie di dati necessari per la successiva generazione delle features, dalle quali si elaborano le stime di ritorno del cliente.

Un'importante raccomandazione prevede che il dataframe sia convertito a matrice prima di procedere con l'elaborazione del modello sui dati. A questa matrice di dati estratti viene applicato il modello di TtNP attraverso il comando predict. Il TtNP appena calcolato viene unito ai dati utilizzati dal modello e vengono calcolate la data dell'ultima spesa e la data della prossima spesa prevista.

Una volta che il Time to Next Purchase è stato calcolato, è possibile acquisire informazioni utili sulla frequenza d'acquisto dei clienti. Ad esempio, se il TtNP ha un valore basso ciò potrebbe mostrare come i clienti tornino ad acquistare più spesso mentre un valore TtNP più alto potrebbe rivelare periodi di inattività più ampi.

4.6.2 Variazione del livello di degrado TtNP

L'indicatore che esprime la variazione del livello di degrado TtNP si basa su un algoritmo che viene riallenato solo nelle volte in cui il cliente torna ad acquistare in negozio (o online), perciò il training set viene modificato attraverso l'inserimento di una nuova osservazione.

Ogni giorno si avranno dei nuovi clienti che vanno a fare la spesa e per essi il TtNP viene ricalcolato sulla base di nuove stime. Inoltre, è possibile calcolare il degrado ovvero la differenza tra il tempo atteso di ritorno del cliente e la recenza (quanto tempo è passato dal suo ultimo acquisto).

Se si calcolano dei valori medi in un opportuno intervallo temporale, si può comprendere il motivo per il quale l'algoritmo non sta esprimendo il comportamento del cliente nel modo corretto. Probabilmente, il problema non è dovuto all'algoritmo ma al fatto che il cliente ha cambiato le proprie abitudini.

In questo caso, gli scenari sono essenzialmente due: il primo prevede che il cliente col passare del tempo vada più frequentemente in negozio mentre il secondo prevede che il cliente vada sempre meno spesso. Rispettivamente, l'intertempo medio verrà dunque (almeno inizialmente) sovrastimato o sottostimato dall'algoritmo.

In entrambi i casi il cliente sta modificando le sue abitudini. Inoltre, il fatto di capire come varia questo output dice in qualche modo come il cliente si sta comportando rispetto alla frequenza delle spese effettuate. In sintesi, questo indicatore è ottenuto tendenzialmente prendendo uno storico di se stesso.

In base al valore della misura TtNP degrado si ottengono diversi cluster di clienti:

1. "Non a rischio";
2. "Da osservare";
3. "Da sollecitare";
4. "Vicino all abbandono";
5. "Churner".

Inoltre, la validità del Time to Next Purchase come indicatore delle abitudini d'acquisto dei clienti può essere influenzato da altri fattori come ad esempio la

stagionalità (per evitare questo problema si può scegliere come orizzonte temporale di analisi l'anno) oppure le campagne di marketing. Dunque, è giusto ricordare che il calcolo del TtNP può essere adattato secondo le specifiche necessità di un'azienda e secondo i dati disponibili.

Ricapitolando, si può affermare che il ciclo (semplificato) che riassume il continuo processo di ricalcolo del TtNP è il seguente:

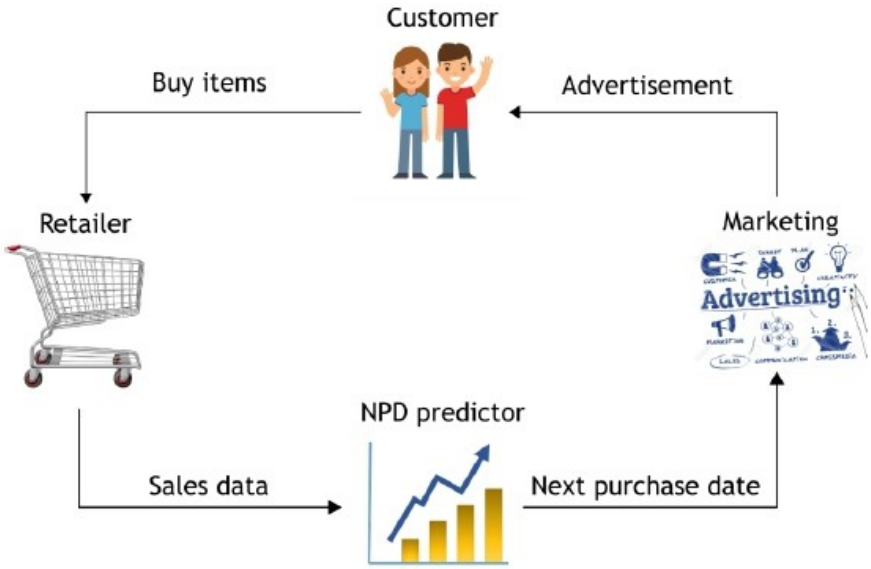


Figura 4.16: TtNP
(37)

Capitolo 5

Progettazione e Sviluppo dell'Acceleratore Data Analytics

In questo capitolo si andrà a progettare e sviluppare l'acceleratore data analytics. Un primo paragrafo sarà dedicato alla descrizione della sua architettura basata su tecnologie cloud, dopodichè verranno analizzate tutte le fasi previste per la sua corretta implementazione.

Esse sono nell'ordine: "raccolta e integrazione dei dati", "trasformazione e preparazione dei dati", "creazione del data warehouse e implementazione della UCV" e infine "creazione di dashboard per l'analisi dei dati retail".

Per eseguire questa serie di passaggi è necessaria una fase di orchestrazione dei flussi dati. Essa consiste in un processo di coordinamento e gestione di svariati sistemi, applicazioni e/o servizi volto a eseguire l'intero flusso di lavoro a cui sono soggetti i dati (in particolare ETL/ELT). In tale situazione, possono esserci molteplici attività (o task) automatizzate che possono coinvolgere più sistemi.

Lo scopo di questo processo è alleggerire e ottimizzare l'esecuzione di processi frequenti e reiterati, assistendo i team nella gestione di flussi di lavoro particolarmente complessi. Nel momento in cui un'attività può essere automatizzata, l'orchestrazione permette di evitare perdite di tempo e quindi di incrementare

l'efficienza e limitare le ridondanze.

Molto spesso, il concetto di orchestrazione viene confuso con quello di automazione. Quest'ultima riguarda perlopiù la programmazione di una singola attività affinché essa venga svolta senza l'aiuto dell'uomo, mentre il software di orchestrazione deve configurare più attività (alcune delle quali possono essere automatizzate) in un unico processo e, inoltre, deve prendere decisioni basate sugli output di un'attività automatizzata per determinare e coordinare le attività successive.

È possibile attuare un processo di orchestrazione su (59):

- Applicazioni: orchestrare le applicazioni significa integrare due o più applicazioni software fra loro. Ciò permette la gestione e il monitoraggio delle integrazioni in modo centralizzato, aggiungendo funzionalità di instradamento dei messaggi, sicurezza, trasformazione e affidabilità;

- Servizi: approccio che fa riferimento all'orchestrazione di microservizi, reti e flussi di lavoro. Ciò permette inoltre il coordinamento e la gestione di sistemi (che non hanno la capacità nativa di integrarsi uno con l'altro) distribuiti fra differenti fornitori e domini in cloud, condizione fondamentale nel mondo di oggi;

- Sicurezza: approccio che permette alle aziende di gestire le minacce automatizzando la loro individuazione e la loro raccolta al fine di ottenere informazioni rilevanti. Per minacce di livello inferiore vengono inoltre identificate le risposte agli incidenti;

- Processi: orchestrare i processi significa integrare strumenti e flussi di lavoro. Più nel dettaglio, con questo termine si fa riferimento alla gestione dell'intero ciclo di vita del processo da un unico punto centralizzato, mediante l'aggregazione di singole attività in processi completi e l'agevolazione delle integrazioni fra sistemi con connettori universali o adattatori di API.

L'avvento del cloud computing (attravero cloud pubblici, privati e ibridi) ha incrementato il livello di complessità e ha reso necessaria l'esigenza di progettare dei software che fossero in grado di gestire e implementare svariate dipendenze su diversi cloud.

Il provisioning (insieme di processi che rende disponibile l'intera infrastruttura IT) dei carichi di lavoro e della capacità di storage dei server è un'attività che solitamente rientra in questo contesto, anche se in alcuni servizi (come ad esempio Cloud Composer di Google) non è prevista.

Altro concetto fondamentale riguarda il cosiddetto livello di orchestrazione. Esso consente in prima battuta di generare connessioni fra il connettore di un'azienda e quelli delle applicazioni di terze parti, in modo da coordinare più servizi API.

In secondo luogo viene gestita la formattazione dei dati fra servizi separati ma viene dato anche supporto alla trasformazione dei dati, alla gestione dei server/autenticazioni e all'integrazione con i sistemi esistenti.

5.1 Architettura dell'acceleratore data analytics basato su tecnologie cloud

Uno degli obiettivi dell'acceleratore data analytics è quello di fornire un'architettura standard basata su tecnologie cloud, che riassuma l'intero processo del dato ("raccolta e integrazione dei dati", "trasformazione e preparazione dei dati", "creazione del data warehouse e implementazione della UCV" e infine "creazione di dashboard per l'analisi dei dati retail"):

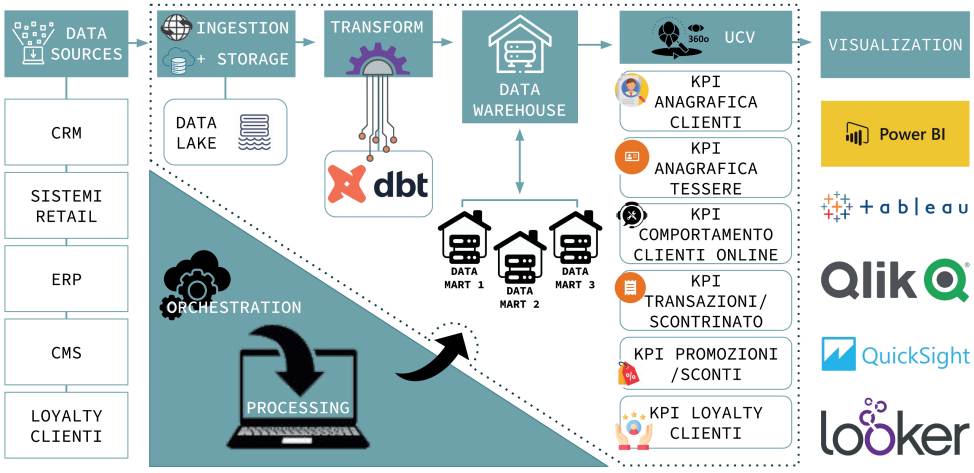


Figura 5.1: Architettura Standard Acceleratore (68)

Durante l'intero capitolo, tale immagine verrà ripresa più volte con l'obiettivo di andare ad evidenziare ogni singola fase prevista dall'acceleratore. Dunque, per ogni paragrafo sarà immediato comprendere quale sia lo step a cui il dato è sottoposto.

5.2 Raccolta e integrazione dei dati

5.2.1 Sistemi di sorgenti dati

La prima fase dell'acceleratore prevede di effettuare la raccolta dati. Per iniziare, è necessario considerare le sorgenti da cui si vogliono estrarre i dati. Esse (riprendendo quanto detto nel capitolo precedente) sono:

1. CRM: software che memorizza sia le informazioni di contatto dei clienti sia le interazioni del cliente stesso con l'azienda, dunque si tratta di un software che ottimizza la gestione di questi rapporti e che di conseguenza dà origine a conversazioni produttive ed efficaci;

2. Sistemi Retail: strumenti che vengono utilizzati per la gestione operativa dei punti vendita. In questa categoria sono inclusi anche i sistemi cassa e gli e-commerce, sorgenti in tempo reale che sfruttano strumenti che sono stati configurati ragionando secondo un concetto di coda.

   Attraverso questa sorgente è possibile analizzare diversi dati tra cui principalmente transazioni dello scontrinato e anagrafiche di clienti, articoli, carte e negozi;

3. ERP: sistema gestionale che integra tutti i processi di business rilevanti di un'azienda (vendite, acquisti, gestione magazzino, contabilità, finanza, etc);

4. CMS: applicativo web che nel contesto del CRM analitico funge da fonte alimentante in grado di descrivere (tramite i dati raccolti sui device e sui social) i comportamenti dei clienti online;

5. Loyalty clienti: fonte dati che tra le varie informazioni raccoglie, per esempio, accumulo e fruizione punti (in cassa e su piattaforme online), donazioni e indicazioni simili.

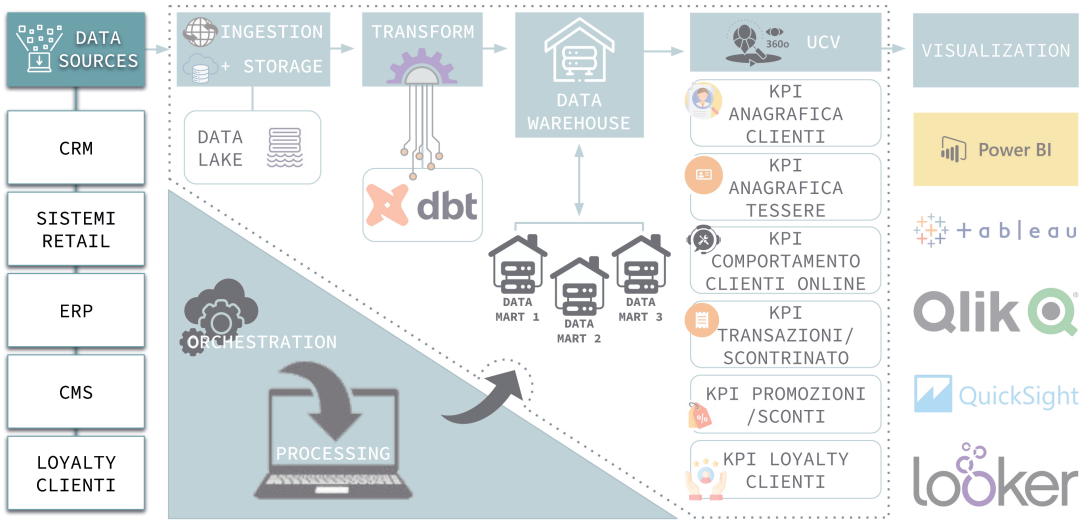


Figura 5.2: Architettura Standard Acceleratore Data Sources (68)

I sistemi sorgente devono poter dialogare con l'acceleratore, quindi a seconda dello scenario (on prem, in cloud o contesto ibrido) potrebbe esserci bisogno di instaurare dei servizi VPN (Virtual Private Network) per il dialogo tra i sistemi di sorgenti dati e il servizio di storage previsto dall'acceleratore. Tale instaurazione consiste in un processo che prevede fasi differenti:

1. Analisi dei requisiti: dopo aver individuato le sorgenti dati da inserire, bisogna definire i requisiti di connettività e sicurezza per assicurare una trasmissione affidabile dei dati;

2. Selezione del tipo di VPN: esistono fondamentalmente due tipi di VPN, o VPN site to site (da gateway a gateway) o VPN client to site (accesso remoto). La differenza tra loro è piuttosto semplice da comprendere.

   Infatti, il primo gestisce connessioni remote tra intere reti mentre il secondo è contraddistinto da connessioni a utente singolo; (48)

3. Configurazione della VPN: durante questa fase bisogna definire le impostazioni di sicurezza (inclusi protocolli di crittografia e autenticazione) e permettere un flusso sicuro dei dati attribuendo indirizzi IP e configurando i tunnel VPN;

4. Autenticazione e autorizzazione: bisogna assicurarsi che esclusivamente gli utenti autorizzati siano in grado di accedere alle risorse. Configurando le autorizzazioni, è possibile accedere solamente ai dati essenziali;

5. Integrazione con i sistemi di sorgenti dati: bisogna garantire sia la connettività VPN (attraverso la configurazione dei sistemi) sia la compatibilità dei protocolli supportati di comunicazione;

6. Monitoraggio e manutenzione: è necessario individuare problemi potenziali di connettività monitorando la performance della VPN e, inoltre, bisogna procedere con aggiornamenti e manutenzione costante per assicurare la dovuta efficienza della connessione;

7. Test e validazione: c'è bisogno di effettuare dei test approfonditi per assicurare uno scambio di dati preciso e puntuale e, successivamente, attraverso la validazione della trasmissione (in contesti reali) è possibile raggiungere il corretto funzionamento della VPN;

8. Documentazione: per semplificare la gestione futura di altre connessioni, è consigliabile documentare tutte le configurazioni, procedure e protocolli attribuiti alla VPN.

5.2.2 Ingestion dei dati

La fase di ingestion è un processo che ha lo scopo di acquisire e importare dati da sorgenti differenti, per il loro immediato impiego oppure per la loro archiviazione in un sistema di gestione dei dati (che solitamente corrisponde al data lake).

Il data lake archivia grossi volumi di dati non filtrati, da usare in un secondo momento per un'obiettivo specifico. Solo nel momento dell'effettiva analisi sono ricavate struttura, integrità, selezione e formato dei differenti set di dati.

Un data lake è l'opzione corretta nel caso in cui un'azienda abbia bisogno di storage a basso costo per dati non formattati e non strutturati (ricevuti da più sorgenti) che si vogliono adoperare per qualche scopo in futuro.

Tra i diversi livelli di storage si riscontra sostanzialmente un livello runtime (che serve per l'analisi immediata dei dati live) e un livello di archiviazione contenente dati pregressi più vecchi di un periodo prefissato (dato storico).

Inoltre, è necessario definire un time life del dato ovvero bisogna applicare delle politiche di retention che impongono la conservazione dei dati, per un arco di tempo che generalmente non deve essere superiore al conseguimento delle finalità per le quali essi sono trattati.

Per rendere il tutto più veloce ed efficiente, si è cominciato ad investire (da un punto di vista operativo) nello sviluppo di software capaci di automatizzare la maggior parte delle attività di data ingestion. Tutto ciò però implica l'esigenza di avere piattaforme abili nell'integrare funzioni di data preparation.

I dati procurati dai sistemi di ingestion possono essere processati in batch o in real time. L'elaborazione batch è un sistema per mettere in pratica grossi volumi di lavoro sui dati di tipo ripetitivo, progettato per essere un processo gestito con un'interazione minima o nulla da parte dell'utente, perciò quasi del tutto automatizzato. L'elaborazione batch è quindi un modo straordinariamente congruo e adatto per elaborare ingenti quantità di dati in poco tempo.

A differenza dell'elaborazione batch, con l'espressione real time analytics si intende un'analisi dei big data che consente di sfruttare tecnologie e processi attraverso i quali i dati devono essere misurati, gestiti e analizzati in tempo reale non appena entrano nel sistema, permettendo visualizzazioni, comprensioni e approfondimenti rapidi alle aziende.

Il passaggio successivo della fase di ingestion è il Data Processing (o elaborazione di dati). Esso racchiude l'intero processo di analisi dei dati grezzi ovvero di tutti quei dati che vengono definiti raw. Questa operazione prevede vari step (17):

1. Raccolta dei dati: essi vengono solitamente salvati in dei data lake, pronti per cominciare il loro processamento;

2. Pulizia dei dati: fase che permette di individuare errori, duplicazioni, inserimenti incompleti o non corretti e quindi ha l'obiettivo di risolvere queste problematiche con l'intento di produrre i migliori dataset possibili per l'elaborazione;

3. Inserimento dei dati: al passo successivo, i dati possono essere introdotti in un sistema di elaborazione attraverso sorgenti di input differenti. Ciò consente di "tradurre" i dati in un formato comprensibile per la soluzione;

4. Elaborazione dei dati: i dati vengono analizzati, raccolti e organizzati in dataset (in base ai propri criteri) grazie all'impiego di algoritmi di AI e ML;

5. Risultati dei dati: le informazioni ottenute possono essere esposte in formati comprensibili come grafici, testi o quant'altro. Tutti gli output del Data Processing possono essere introdotti nuovamente nel sistema per creare elaborazioni sempre aggiornate e significative;

6. Storage dei dati: l'ultima fase prevede il trasferimento dei dati su uno storage (a basso costo e stabile) a blocchi, che può essere Cloud Storage (su GCP), S3 (su AWS), Block Storage (su Azure), htfs (su cluster hadoop) o glusterfs (su kubernetes). Le informazioni salvate verranno poi sfruttate per le elaborazioni successive.

Riprendendo quindi l'architettura mostrata ad inizio capitolo, in questo momento ci si trova nella fase seguente:

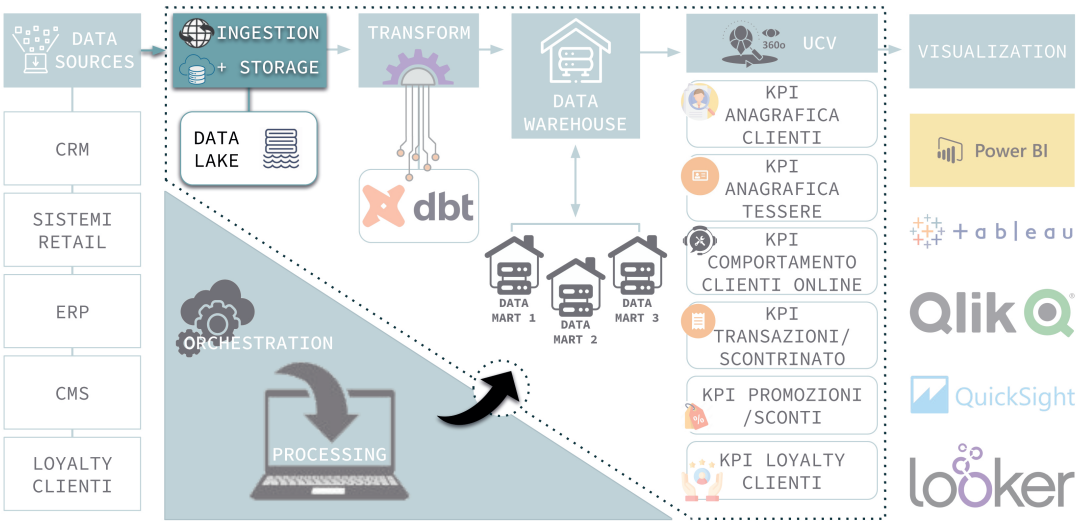


Figura 5.3: Architettura Standard Acceleratore Ingestion + Storage (68)

Nella fase di storage entra in gioco il GDPR (Regolamento Generale sulla Protezione dei Dati), che non prevede soltanto che i dati siano archiviati nel rispetto delle normative locali ma conferma anche alle autorità competenti che gli stessi dati sono protetti e conformi in ogni circostanza.

Inoltre, viene previsto l'utilizzo del KMS, un sistema integrato di sicurezza ideato per la gestione (quindi creazione, distribuzione e manutenzione) delle chiavi crittografiche che sono impiegate per la protezione dei dati ritenuti sensibili. In tal modo, si cerca di non compromettere la loro sicurezza andando a prevenenire tutti gli accessi non autorizzati.

La sicurezza può essere inoltre ottimizzata attraverso un processo regolare di rotazione delle chiavi. Il KMS semplifica questa procedura mediante la gestione di nuove key e l'aggiornamento nei sistemi (i quali devono essere perfettamente integrati).

5.3 Trasformazione e preparazione dei dati

Dopo la raccolta e l'integrazione dei dati, la fase successiva consiste nella trasformazione e preparazione dei dati, la quale prevede diverse azioni che possono essere il filtraggio, l'ordinamento, la pulizia o l'aggregazione. Tali azioni portano alla creazione di 4 livelli del dato (raw, clean, curated e publicated) che permettono in seguito di creare il data warehouse e i relativi data mart.

Lo strumento (open-source) che permette all'acceleratore di essere agevole e performante nella fase di trasformazione (e quindi che dà una marcia in più alla soluzione fornita) è dbt, anche detto data build tool.

Esso consiste in un flusso di lavoro che permette ai team di distribuire in modo rapido e collaborativo il codice di analisi seguendo le migliori pratiche di ingegneria del software come modularità, portabilità, CI/CD (Continuous integration/Continuous Delivery) e documentazione.

Di seguito un'immagine generalista ma esplicativa di come lavora lo strumento:

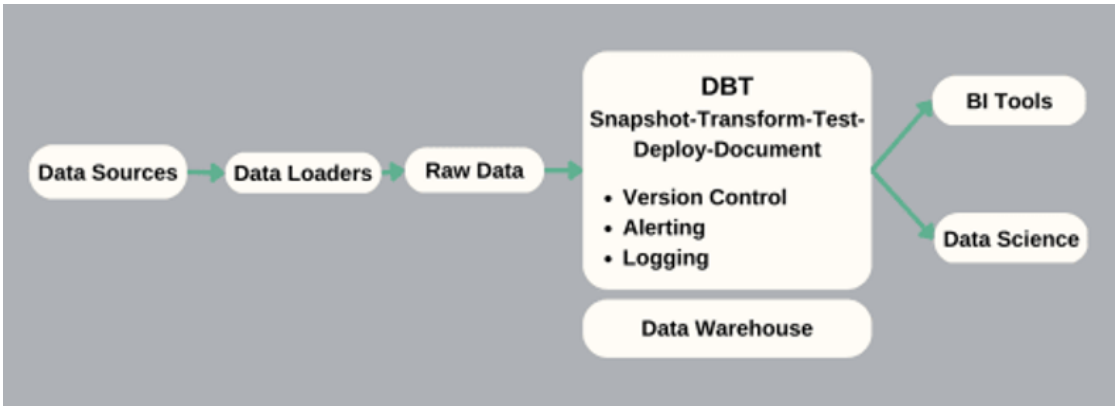


Figura 5.4: dbt come lavora
(68)

5.3.1 Vantaggi dello strumento dbt

Ci sono diversi tool di trasformazione del dato, ma i vantaggi di dbt che verranno descritti e analizzati in questo paragrafo evidenziano al meglio le potenzialità che nessun altro o pochi strumenti possono offrire. I motivi per i quali è possibile considerare dbt come punto centrale della soluzione fornita dall'acceleratore sono i seguenti:

- Ottimizzazione dei costi di implementazione di una data platform;
- Non è necessario scrivere il codice per creare tabelle e viste. dbt integra le istruzioni select all'interno del codice che si definisce;
- Consente sia di utilizzare strutture di controllo nelle query sia di condividere SQL ripetuti tramite macro, utilizzando linguaggi di template leggeri;
- Ottenimento del controllo della versione con Git, software open-source che permette di registrare ogni modifica che viene attuata ai codici, di confrontare versioni differenti di uno stesso file e infine di cooperare simultaneamente a uno stesso programma per implementare nuove funzionalità (senza interferire con il lavoro degli altri membri del team);

- Documentazione dei modelli generata in modo automatico (ottimizzazione dei tempi);

- Test e suggerimenti per migliorare l'integrità di SQL.

Successivamente, si passa alle modalità di utilizzo dello strumento, tra le quali se ne possono riscontrare sostanzialmente due:

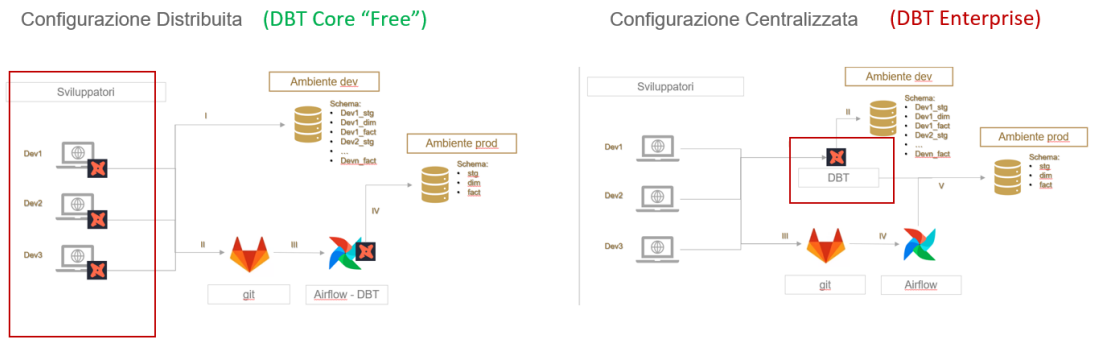


Figura 5.5: dbt modalità utilizzo (68)

In entrambe le configurazioni è presente Git, sistema di controllo della versione distribuita che gestisce versioni multiple tenenti traccia di ogni cambiamento e modifica applicata al codice.

Gli sviluppatori effettuano il commit ossia lo snapshot di tutti i file in un momento specifico del lavoro in locale, sincronizzando la copia del repository con la copia nel server. Dunque, i client non sono obbligati a sincronizzare il codice con un server prima di poter dare origine a nuove versioni del codice (ciò accade invece con il controllo della versione centralizzato). (58)

Per quanto riguarda Airflow–dbt, esso consiste in uno schedulatore che consente di effettuare automaticamente i rilasci in produzione. Rispetto ad Airflow che è progettato per concentrarsi maggiormente sul flusso effettivo di dati nella sua interfaccia, dbt offre un'interfaccia front-end completa per lo sviluppo e la codifica delle query.

5.3.2 Livelli di trasformazione

Durante la fase di trasformazione, i vari layer all'interno del data warehouse devono contenere rispettivamente 4 livelli: nel livello 1 sono contenuti i dati raw, nel livello 2 e 3 sono contenuti i dati clean e curated mentre nel livello 4 sono contenuti i dati publicated:

Figura 5.6: Architettura Standard Trasformazione (68)

Tornando a dbt, altre sue caratteristiche riguardano prevalentemente la definizione di modelli. Essi corrispondono a rappresentazioni logiche di una tabella o di una vista che sono definite in file ".sql". Le dipendenze dei modelli vengono gestite in modo automatico e ottimale per assicurare un loro preciso ordine di esecuzione.

In aggiunta, bisogna anche ricordare che lo strumento dbt è in grado di generare documentazione basata sul codice, il che permette di comprendere in modo semplice l'obiettivo dei modelli e la loro gerarchia. Inoltre, sono supportati snapshotting e versioning con i quali è possibile tenere traccia delle modifiche dei dati e dei loro cambiamenti nel tempo.

Infine, adoperando un semplice connettore di dbt, è possibile supportare diverse Data Platforms tra cui quelle ufficiali come ad esempio Azure Synapse, BigQuery e Databricks.

Ecco dunque che con queste ultime considerazioni termina la fase di trasformazione e preparazione dei dati.

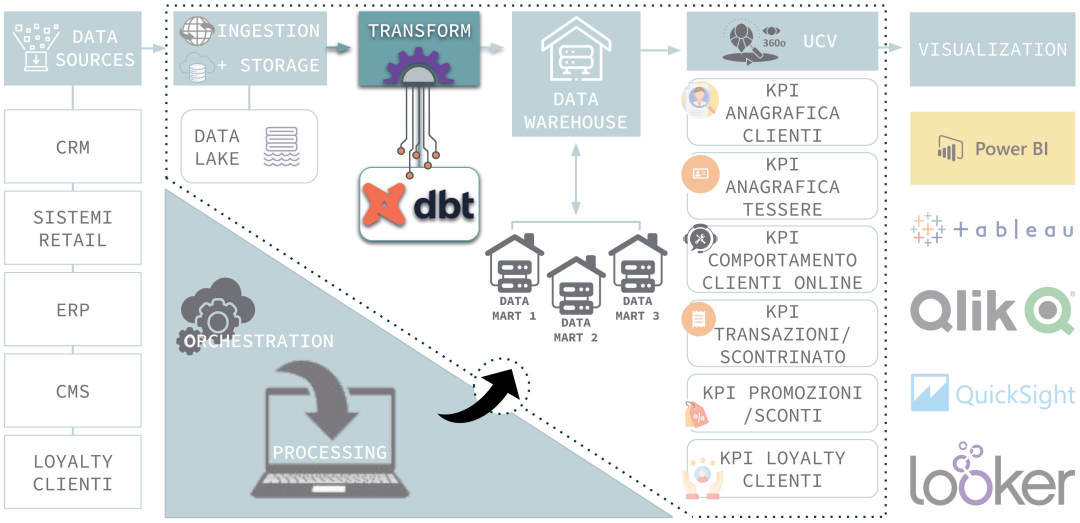


Figura 5.7: Architettura Standard Acceleratore Transform (68)

5.4 Creazione del data warehouse e implementazione della UCV

Dopo aver affrontato le prime fasi previste dall'acceleratore (raccolta, ingestion e trasformazione dei dati), è il momento di proseguire con la seconda parte caratterizzante l'architettura andando prima ad analizzare il paragrafo relativo alla creazione del data warehouse (con i rispettivi data mart) e successivamente il paragrafo relativo all'implementazione della UCV (con i rispettivi KPI).

5.4.1 Architettura DWH e data mart

Il data warehouse è la base di dati che ha l'obiettivo di dare supporto alle decisioni, ed è mantenuta in modo separato dalle basi di dati operative dell'azienda. I dati che la caratterizzano sono orientati ai soggetti di interesse, integrati, consistenti,

dipendenti dal tempo, non volatili e impiegati per dare sostegno alle decisioni aziendali.

I dati sono separati sia per una questione di performance (ricerche complesse limitano fortemente le prestazioni delle transazioni operative) sia per una questione di gestione (informazioni mancanti, consolidamento e qualità dei dati).

Al data warehouse viene spesso associato il concetto di data mart, un suo sottoinsieme focalizzato su contesti business specifici (che può essere alimentato dal DWH primario oppure direttamente dalle sorgenti). La realizzazione richiede meno tempo ma esige una progettazione attenta, in modo da impedire più avanti problemi di data integration.

L'alimentazione del data warehouse avviene (come ribadito più volte) dal processo ETL, il quale prevede l'estrazione dei dati da sorgenti esterne, la pulizia dei dati (errori, dati mancanti o duplicati), la trasformazione/conversione di formato e infine il caricamento e refresh periodico.

Generalmente, l'architettura del data warehouse può essere di due tipi: a 2 livelli o a 3 livelli. Le caratteristiche dell'architettura a 2 livelli prevedono una certa semplicità nel suddividere il carico transazionale (OLTP) da quello analitico (OLAP) ma dall'altro lato c'è il bisogno di compiere "al volo" la preparazione dei dati (ETL).

Invece, nell'architettura a 3 livelli viene inserito un livello di alimentazione che corrisponde alla staging area, un'area di transito che consente di suddividere l'elaborazione ET dal caricamento nel data warehouse. In aggiunta, essa consente operazioni complesse di pulizia/trasformazione dei dati e propone inoltre un modello integrato dei dati aziendali. L'aspetto però negativo è dovuto alla ridondanza addizionale data dal maggior spazio richiesto per i dati.

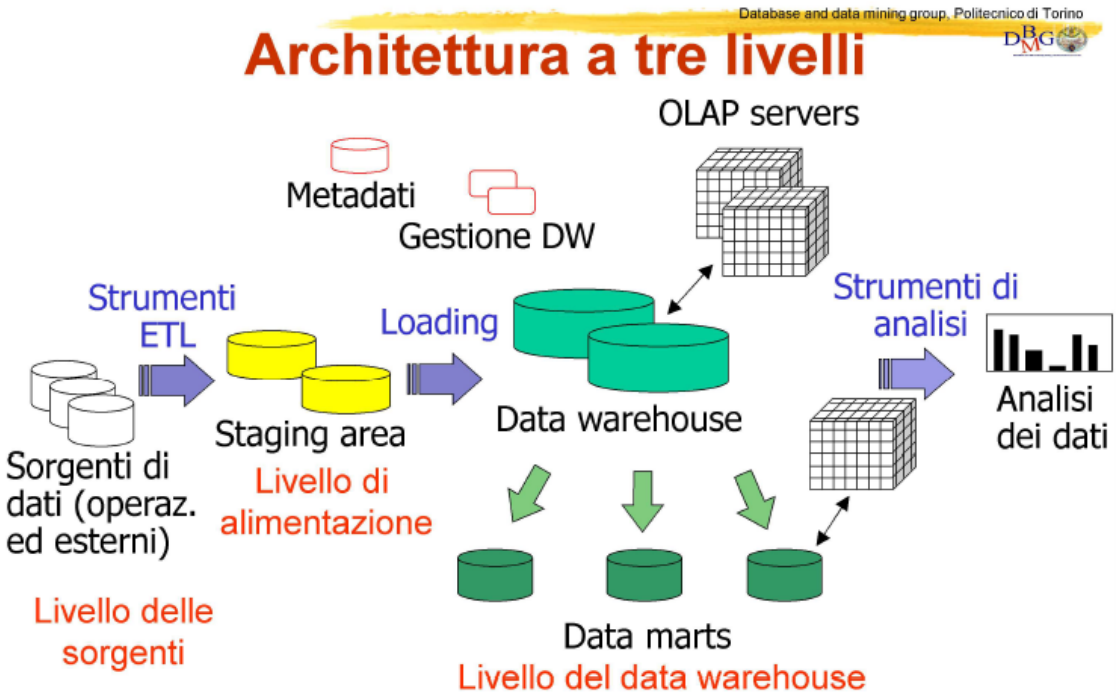


Figura 5.8: Architettura a 3 livelli DWH (62)

Con questo tipo di architettura, il processo ETL viene semplificato dalla presenza di una staging area e viene eseguito durante il primo popolamento del DWH e l'aggiornamento periodico dei dati.

Invece, per quanto riguarda la progettazione del data warehouse, esistono due tipologie di approcci:

- Approccio top-down: consiste in una realizzazione che procura una visione globale e completa dei dati aziendali. Tuttavia, l'implementazione richiede un costo significativo, un tempo di realizzazione prolungato e delle analisi complesse.

- Approccio bottom-up: consiste in una realizzazione incrementale che avviene attraverso l'inserimento di data mart focalizzati su contesti di business specifici. A differenza dell'approccio top-down, il costo e il tempo di consegna sono entrambi contenuti.

Con ciò termina la fase relativa alla creazione del DWH dunque considerando l'architettura proposta dall'acceleratore, il tema appena descritto si trova esattamente nella zona messa in risalto nell'immagine sottostante:

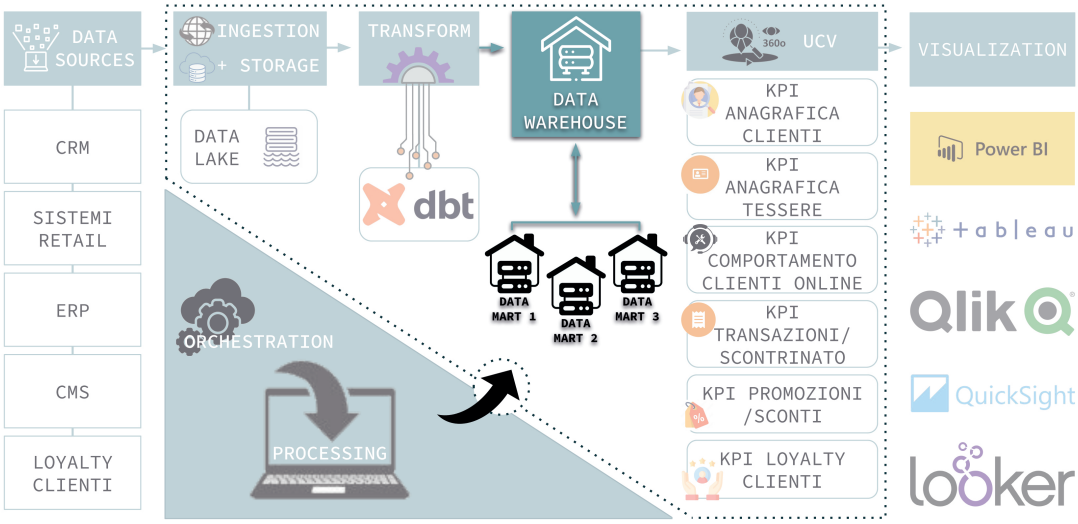


Figura 5.9: Architettura Standard Acceleratore Data Warehouse (68)

5.4.2 Costruzione degli indicatori chiave della UCV

Dopo la creazione del data warehouse, è il momento di concentrare l'attenzione sull'implementazione della UCV, la quale deve essere arricchita con i KPI analizzati nel capitolo precedente. Essa rappresenta quel luogo univoco centralizzato che permette di archiviare ogni informazione sui clienti in maniera aggregata, coerente e olistica.

Tutto ciò consente di avere una panoramica puntuale sugli utenti stessi, con la quale si è in grado di visionare ogni loro tipo di movimento e interazione indipendentemente dai canali utilizzati (negozio, sito web, app, social, e-commerce, etc).

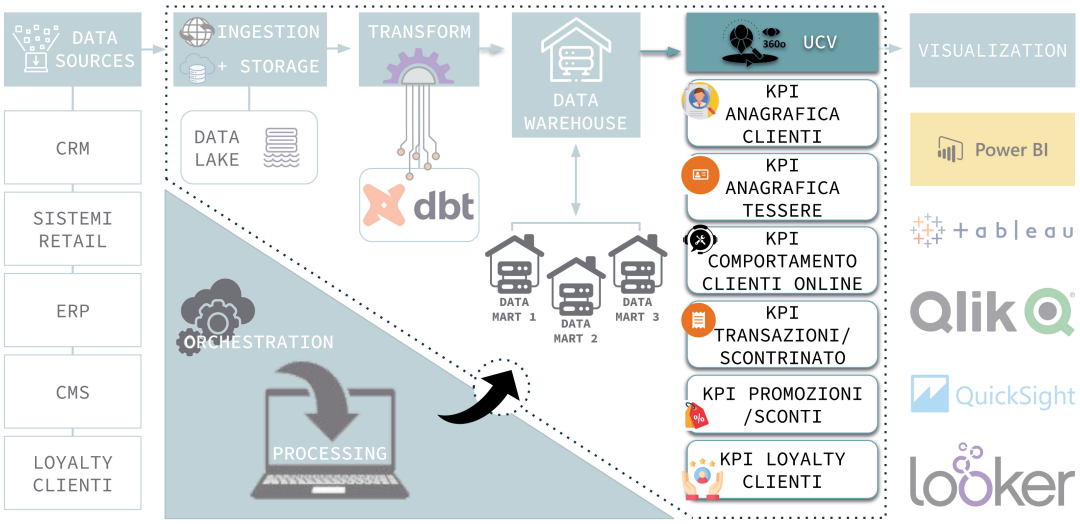


Figura 5.10: Architettura Standard Acceleratore UCV (68)

I vantaggi offerti dalla UCV sono molteplici. Tra i pricipali rientrano sicuramente una targetizzazione più precisa, una customer experience migliore (si propone al cliente esattamente ciò che desidera e senza ritardi), un'ottimizzazione dei processi e della comunicazione interna (è possibile procurarsi più facilmente le informazioni necessarie su un utente, senza dover contattare altri reparti) e una facile individuazione del customer journey. (84)

Dunque, tramite dbt vengono implementate le query di estrazioni dati e creati data mart tematici propedeutici per la realizzazione della UCV. Ecco un riassunto degli indicatori standard più importanti:

- KPI inerenti all'anagrafica clienti: data partizione, numero cliente, codice fiscale, nome, cognome, indirizzi relativi a domicilio/residenza, data di nascita/età, codice negozio radicamento, canale, data inizio relazione, anzianità cliente, email, cellulare, opt in 1, opt in 2, opt in 3, mezzo contatto preferito;
- KPI inerenti all'anagrafica tessere: tipologia di carta;
- KPI inerenti al comportamento dei clienti online: data creazione/modifica profilo, anzianità, login web/app, recenza (online);

- KPI inerenti alle transazioni/scontrinato: numero carta più utilizzata, frequenza negozio radicamento, canale d'acquisto prevalente, spesa per canale, id negozio spesa/frequenza prevalente, giorno acquisto/fascia oraria prevalente, recenza, frequenza, intertempo medio acquisti, valore monetario, scontrino medio, recenza meno intertempo medio ultimi acquisti, spesa/frequenza media mensile rolling, valore cliente, reparto misure (ad esempio la frequenza/spesa per reparto);

- KPI inerenti alle promozioni e agli sconti: percentuale promo valore, percentuale promo quantità, prodotti sempre in promozione, cassaforte sconto, valore risparmio, valore risparmio year to date, score "nome reparto", indice di Gini;

- KPI inerenti al comportamento dei clienti rispetto a programmi di Loyalty ed iniziative commerciali e di marketing: raccolta dei punti.

Con questa serie di indicatori relativi alla customer analytics, la UCV è finalmente pronta per dare il proprio supporto alla creazione di dashboard.

5.5 Creazione di dashboard per l'analisi dei dati retail

L'intero processo a cui deve far fronte il dato retail termina dunque con la creazione delle dashboard, strumenti visivi che si focalizzano sull'evidenziare i dati attraverso i KPI più importanti dei clienti (che sono stati costruiti e inseriti all'interno della UCV). Mediante il loro utilizzo sarà possibile dare sostegno e appoggio alle decisioni che dovranno essere prese dall'azienda, per mantenere e consolidare in primis la customer base relativa ai propri clienti.

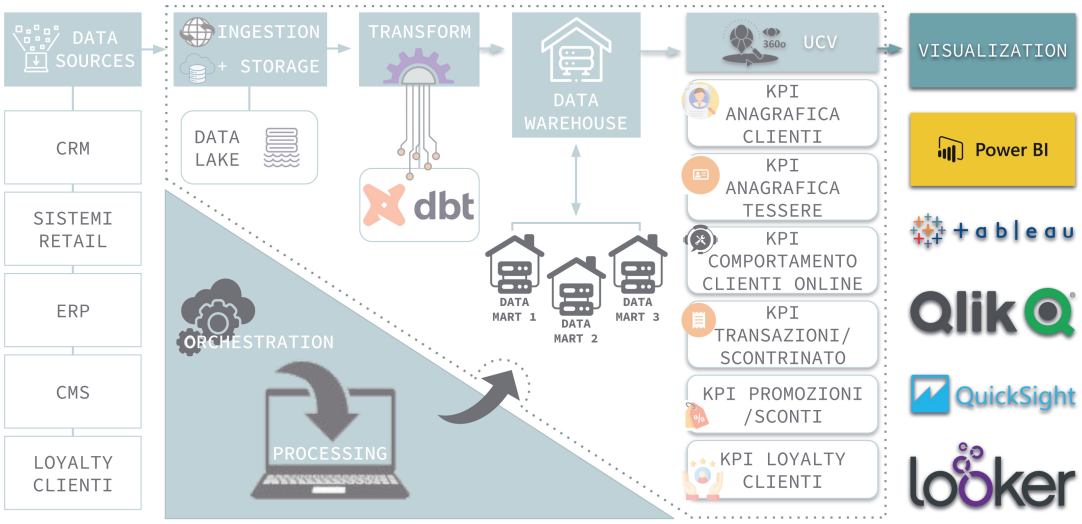


Figura 5.11: Architettura Standard Acceleratore Visualization (68)

5.5.1 Vantaggi offerti dalle dashboard

Una dashboard è in grado di offrire molteplici vantaggi solo nel caso in cui essa venga ideata, progettata (basandosi esclusivamente sui dati) e impiegata efficacemente. I benefici principali che si possono riscontrare sono i seguenti (82):

- Aggiornamenti in real time: in questo caso è possibile prendere decisioni aziendali efficaci e intelligenti in ogni momento. Tali aggiornamenti sono inoltre proficui per agevolare le relazioni periodiche (su mese, trimestre, quadrimestre, semestre o anno);

- Risparmio di tempo: non c'è più la preoccupazione di dover recuperare i dati tra molteplici fonti (considerando che si trovano tutti in UCV), perciò si risparmia tempo ed energia rendendo più efficente l'intero processo;

- Maggiore visibilità e trasparenza: vengono procurate informazioni rilevanti sui KPI affinchè sia possibile individuare potenziali trend (positivi o negativi). Per evitare di perdersi in mezzo ai troppi numeri, le dashboard provano ad

attirare l'attenzione su metriche specifiche, al fine di potersi dedicare su ciò che conta di più per l'azienda;

- Maggiore comunicazione: i membri del team possono sostenere una comunicazione duratura e costante grazie alla natura in tempo reale dello strumento;

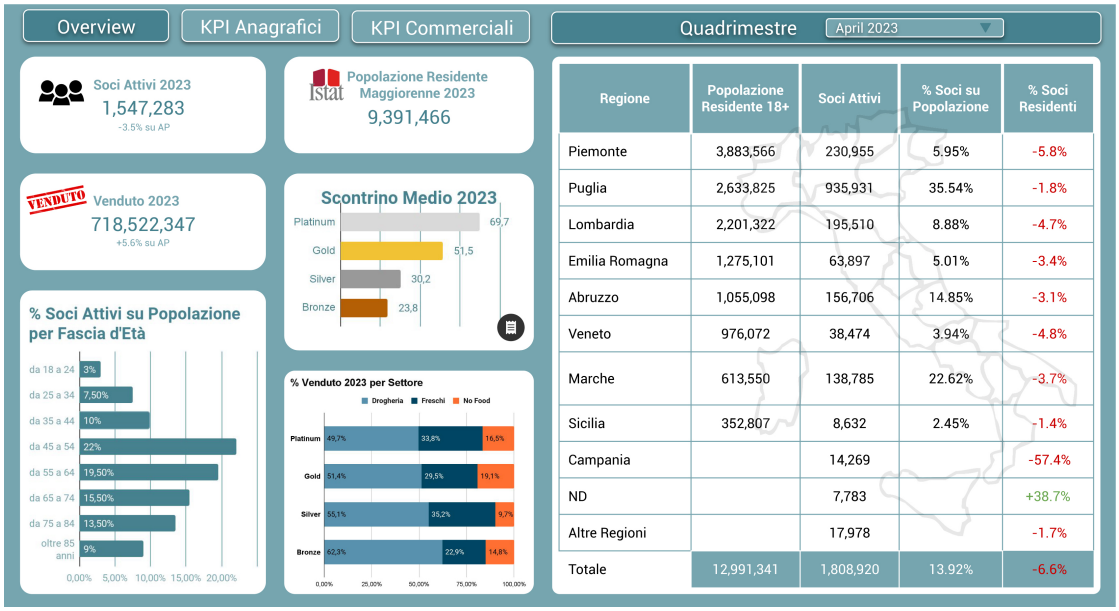
- Maggiore responsabilità: i membri del team possono contare sulla fiducia reciproca del gruppo e sulle proprie responsabilità. Infatti, le dashboard sono in grado sia di rendere i dati costantemente disposibili a tutte le parti coinvolte sia di contribuire a stimolare consegne di lavoro puntuali ed efficienti.

5.5.2 Esempio di dashboard

Una dashboard può essere costruita in svariati modi in base alle specifiche necessità e i diversi ambiti. Solitamente esse vengono popolate attraverso l'inserimento di KPI affini al contesto (in questo caso la customer analytics per il mondo Retail) o mediante l'aggiunta di grafici statistici (istogrammi/grafici a barre, grafici a linee, grafici a torta, diagrammi cartesiani/grafici di dispersione, cartogrammi, box plot, heatmap, etc).

Prima dei prototipi e delle dashboard vere e proprie è utile affidarsi ai cosiddetti mockup, anteprime grafiche particolarmente utili che consentono sia di far capire all'utente come potrebbe essere il risultato finale sia di avere un confronto con il cliente per vedere se l'output rispecchia le sue aspettative oppure se ci sono modifiche da attuare. (15)

Ecco un esempio di dashboard contenente KPI e statistiche che fanno riferimento alla customer analytics:



Regione	Popolazione Residente 18+	Soci Attivi	% Soci su Popolazione	% Soci Residenti
Piemonte	3,883,566	230,955	5.95%	-5.8%
Puglia	2,633,825	935,931	35.54%	-1.8%
Lombardia	2,201,322	195,510	8.88%	-4.7%
Emilia Romagna	1,275,101	63,897	5.01%	-3.4%
Abruzzo	1,055,098	156,706	14.85%	-3.1%
Veneto	976,072	38,474	3.94%	-4.8%
Marche	613,550	138,785	22.62%	-3.7%
Sicilia	352,807	8,632	2.45%	-1.4%
Campania		14,269		-57.4%
ND		7,783		+38.7%
Altre Regioni		17,978		-1.7%
Totale	12,991,341	1,808,920	13.92%	-6.6%

Figura 5.12: Esempio Dashboard (68)

Capitolo 6

Applicazioni casi d'uso

L'obiettivo di quest'ultimo capitolo è quello di mostrare come l'acceleratore possa funzionare in applicazioni di casi d'uso che possono risultare differenti sia a livello di tipologia di cliente coinvolta (contesto GDO, GDS, fashion, beauty) sia a livello di tecnologia impiegata.

In tal caso, lo scopo prefissato è quello di adattare l'architettura proposta dall'acceleratore prima su una soluzione GDO basata su Google Cloud e successivamente su una soluzione fashion basata su Amazon Web Services.

L'intento è quello di mostrare la facilità di adattamento dell'acceleratore data analytics, a prescindere dal contesto retail e a prescindere dalla tecnologia cloud utilizzata (Google Cloud e AWS rientrano, insieme a Microsoft Azure, tra i principali cloud provider sul mercato).

6.1 Scenario 1: soluzione GDO basata su Google Cloud

6.1.1 Contesto scenario 1

Partendo dal primo scenario, si andrà a presentare ed esporre uno use case in contesto GDO basato sulla tecnologia Google Cloud. Andando a modificare l'architettura generale in base al contesto appena descritto, ogni singola fase prevista

dall'acceleratore viene svolta grazie a strumenti focalizzati su componentistiche Google:

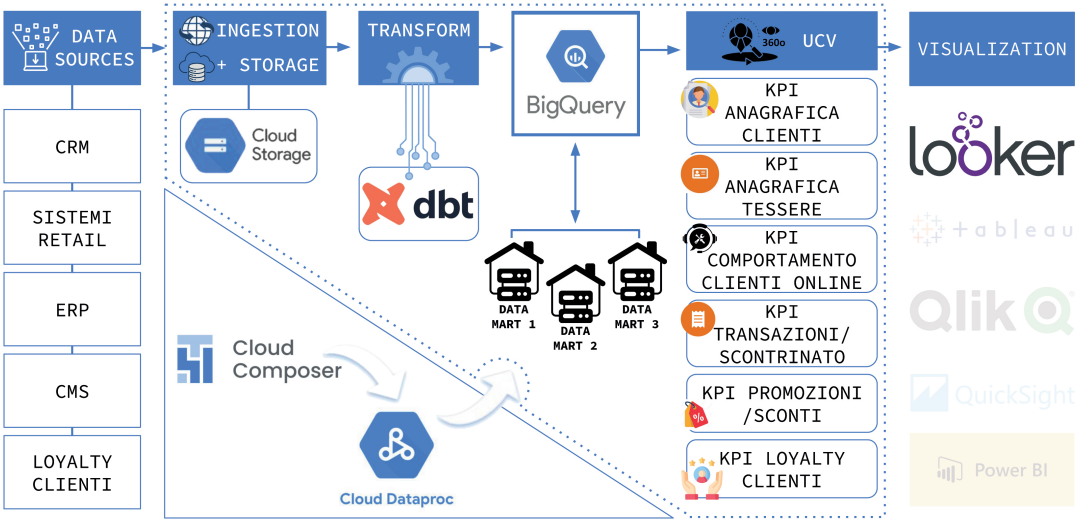


Figura 6.1: Architettura scenario 1
(68)

6.1.2 Obiettivi e strumenti tecnologici scenario 1

Questo tipo di scenario punta ad avere un CRM analitico che sia in grado di condurre e supportare analisi sulla customer base mirate a:

- Limitare il Churn Rate ovvero il tasso di abbandono del cliente;
- Acquisire nuovi clienti tramite incentivi all'acquisto (per esempio attraverso sconti e promozioni);
- Aumentare il venduto mediante l'analisi della UCV, le campagne marketing, il volantino personalizzato e le newsletter.

Per raggiungere gli obiettivi prefissati, è necessario descrivere ogni strumento impiegato nelle varie fasi dell'architettura Google (68):

- Cloud Composer: servizio di orchestrazione del flusso di lavoro completamente gestito. Piuttosto che focalizzarsi sul provisioning delle risorse, esso permette

facilmente di creare, pianificare e monitorare le pipeline a prescindere dal contesto (on-prem , in cloud, in ambienti ibridi o multi-cloud) creando così un ambiente dati unificato.

L'integrazione end-to-end con diversi prodotti Google Cloud (Pub/Sub, Cloud Storage, Dataflow, Dataproc, BigQuery e AI Platform) assicura agli utenti la libertà di orchestrare completamente le loro pipeline.

Cloud Composer fa riferimento ad Apache Airflow, progetto open source che svincola i clienti dalla dipendenza dal fornitore, procura una certa integrazione attraverso un'ampia gamma di piattaforme e offre portabilità.

Attraverso il linguaggio di programmazione Python, le pipeline di Cloud Composer sono configurate come grafi diretti aciclici (DAG), dunque sono di facile impiego per qualsiasi utente a prescindere dal relativo livello d'esperienza.

Inoltre, la risoluzione dei problemi viene gestita meglio grazie sia ad un accesso semplice e immediato ad una ricca libreria di connettori sia a svariate rappresentazioni grafiche del flusso di lavoro in azione. La sincronizzazione automatica dei DAG assicura la puntualità dei job rispetto alla programmazione.

Ultima ma non meno importante è la questione prezzi. In tal caso, lo strumento Cloud Composer impiega un modello di prezzi a consumo, dunque si paga in base all'utilizzo (misurato in termini di vCPU l'ora e GB trasferiti al mese). (14)

- Data Proc: servizio gestito che permette di trarre profitto da strumenti di dati open source per l'elaborazione batch, le query, il trasferimento di flussi di dati e il ML.

  L'automazione di Dataproc permette di dare origine a cluster in maniera spedita, gestirli con semplicità e risparmiare disattivando i cluster stessi nel momento in cui essi non servono.

  L'automazione dello strumento consente di focalizzare il lavoro dei data analyst sui job e sui dati, evitando di perdere tempo e denaro per l'amministrazione e la gestione.

  Lo strumento Dataproc mette a disposizione diversi vantaggi e benefici da cui poter trarre profitto (19):

1. Costo ridotto: Dataproc addebita esclusivamente ciò che viene utilizzato, con una fatturazione (calcolata su CPU virtuale nel cluster all'ora) piuttosto contenuta. Inoltre, le istanze prerilasciabili introdotte dai cluster limitano ancor più i costi;
2. Rapido: operazioni quali avvio, scalabilità e arresto dei cluster Dataproc impiegano in media 90 secondi o meno. Invece, la creazione di cluster on-prem o attraverso provider IaaS può richiedere dai cinque minuti fino alla mezz'ora di tempo in alcuni casi. Questo consente agli analisti e tutti i membri del team di avere più tempo diretto per lavorare sui dati piuttosto che incorrere nell'attesa di avere a disposizione gli stessi cluster;
3. Integrato: viene concessa non solo la disposizione di un cluster ma anche una piattaforma dati completa. Ciò è possibile grazie all'integrazione dello strumento Dataproc con altri servizi Google Cloud Platform (ad esempio Cloud Storage, Cloud Bigtable, BigQuery, etc);
4. Gestito: l'interazione con lo strumento avviene in maniera semplice grazie alla console Google Cloud, a Cloud SDK o all'API REST Dataproc (non è necessaria l'assistenza di un amministratore/software particolare). Non bisogna preoccuparsi di perdere dati poichè, come evidenziato nel punto precedente, lo strumento Dataproc è integrato con altri servizi di storage;
5. Semplice: è possibile trasferire in maniera semplice i progetti esistenti senza dover ridefinire lo sviluppo. Ciò si verifica a causa del fatto che non c'è bisogno di imparare a usare nuovi strumenti o API per impiegare questo tipo di strumento.

- Cloud Storage: è un servizio impiegato nella fase di ingestion + storage, gestito per archiviare dati non strutturati sottoforma di oggetti. Quest'ultimi corrispondono a dati immutabili costituiti da file di qualsiasi formato, che vengono archiviati in contenitori chiamati bucket.

  Tutti i bucket sono associati a un progetto e ogni progetto può essere raggruppato in un'organizzazione. Oggetti, bucket e progetti costituiscono delle risorse in Google Cloud. Dunque dopo aver creato un progetto, si crea un bucket Cloud Storage all'interno del quale è possibile caricare e scaricare gli oggetti.

Cloud Storage dispone di 4 classi di archiviazione principali:

1. Standard Storage: è considerata la migliore per i dati ad accesso frequente o "hot", dunque in generale è ottima per i dati archiviati per brevi periodi di tempo;
2. Nearline Storage: questa soluzione è ideale per archiviare dati a cui si accede raramente (una volta al mese o meno, in media) sia in lettura sia in scrittura. Gli esempi includono backup di dati, contenuti multimediali a coda lunga o archiviazione di dati;
3. Coldline Storage: ulteriore alternativa a basso costo per l'archiviazione dei dati a cui si accede raramente. Tuttavia, rispetto alla classe di archiviazione Nearline Storage, Coldline Storage è pensata per leggere o modificare i dati una volta ogni 90 giorni al massimo;
4. Archive Storage: si tratta della scelta più economica, impiegata idealmente per l'archiviazione dei dati, il backup online e il ripristino di emergenza. È l'opzione migliore per i dati a cui si prevede l'accesso meno di una volta all'anno, perché ha costi maggiori per l'accesso e l'operatività dei dati e una durata minima di archiviazione di 365 giorni.

- BigQuery: è un data warehouse serverless completamente gestito. Ciò significa che BigQuery stesso si prende cura nel procurare risorse e gestire i server nel backend, quindi in generale si prende cura dell'intera infrastruttura sottostante.

  In questo modo, piuttosto che preoccuparsi di distribuzione, scalabilità e sicurezza, l'utente (insieme ai suoi colleghi di team) può concentrare il proprio lavoro sull'impiego delle query SQL per rispondere alle domande aziendali nel frontend.

  BigQuery fornisce due servizi in uno: archiviazione e analisi. È un luogo in cui è possibile archiviare petabyte di dati (per dare un'idea, 1 petabyte corrisponde a 11.000 film in qualità 4K) ma è anche un luogo in cui analizzare i dati con funzionalità integrate come ML e BI.

  Su BigQuery il pagamento avviene sia in base al consumo di byte di dati elaborati dalle diverse query sia in base all'eventuale spazio di archiviazione permanente previsto per le tabelle.

Senza che sia richiesta alcuna azione da parte del cliente, i dati in BigQuery vengono crittografati quando sono inattivi per impostazione predefinita. Per crittografia dei dati inattivi si intende la crittografia impiegata per la protezione dei dati archiviati su un disco.

BigQuery possiede funzionalità di ML integrate che consentono la scrittura di modelli sfruttando il linguaggio SQL. Al termine di una pipeline di dati, lo scopo del motore di analisi di BigQuery è l'acquisizione di tutti i dati elaborati dopo il processo ETL, la loro archiviazione, la loro analisi ed eventualmente un loro ulteriore utilizzo sottoforma di output.

Tali output possono essere fondamentalmente di due tipologie: strumenti di business intelligence (ad esempio Looker) e strumenti di intelligenza artificiale/machine learning. Il primo caso coinvolge figure professionali quali analisti aziendali e analisti dati mentre il secondo caso coinvolge più gli ingegneri di ML e i data scientist (questi strumenti AI/ML fanno parte di Vertex AI, la piattaforma di machine learning unificata di Google).

BigQuery è come un'area di gestione temporanea comune per i carichi di lavoro di analisi dei dati. I servizi di archiviazione e analisi sono collegati da una rete interna ad alta velocità di Google, che permette a BigQuery di scalare le due entità in modo indipendente in base alla domanda.

BigQuery è in grado di importare set di dati da svariate sorgenti inclusi dati interni (dati salvati direttamente in BigQuery), dati esterni, dati multi-cloud (dati archiviati in più servizi cloud, come AWS o Azure) e set di dati pubblici.

Una volta archiviati in BigQuery, i dati vengono completamente gestiti e vengono replicati, sottoposti a backup e configurati per la scalabilità in maniera del tutto automatica.

BigQuery dà anche la possibilità di effettuare query su origini dati esterne, come i dati archiviati su Cloud Storage o su altri servizi di database come Spanner e Cloud SQL.

Ciò significa che un file CSV non elaborato in Cloud Storage o un foglio Google può essere impiegato per la scrittura di una query senza essere prima importato da BigQuery.

Esistono tre modelli di base per caricare i dati in BigQuery:

1. Caricamento batch: i dati di origine vengono caricati in una tabella BigQuery in un'unica operazione batch. Questa può essere un'operazione una tantum o può essere automatizzata in modo che venga eseguita in base a una pianificazione. Un'operazione di caricamento batch può dare origine ad una nuova tabella o inserire dati ad una tabella già esistente;
2. Streaming: batch di dati più contenuti vengono trasmessi in modo continuo affinchè i dati possano essere disponibili per l'interrogazione quasi in tempo reale;
3. Dati generati: le istruzioni SQL vengono impiegate per l'inserimento di righe in una tabella già esistente oppure per la scrittura dei risultati di una query in una tabella.

- Looker: strumento di BI che supporta BigQuery e decine di database SQL differenti. Esso permette agli sviluppatori di definire un livello di modellazione semantica sopra i database impiegando Looker Modeling Language o LookML.

  Quest'ultimo definisce logica e permessi indipendenti da un particolare database o da un linguaggio SQL, il che libera un data engineer dall'interazione con i singoli database e permette di concentrare maggiormente le attenzioni sulla logica aziendale all'interno di un'organizzazione.

  La piattaforma Looker è basata al 100% sul Web, il che facilita l'integrazione nei flussi di lavoro esistenti e la condivisione con più team di un'organizzazione. Esiste anche un'API Looker, che può essere impiegata per incorporare i report Looker in altre applicazioni.

  In base alle metriche più rilevanti per l'azienda, è possibile dare origine a dashboard Looker che procurano semplici presentazioni al fine di sostenere tutti i membri del team nella visualizzazione immediata di uno stato aziendale ad alto livello. Looker propone diverse alternative di visualizzazione ovvero grafici ad area, grafici a linee, diagrammi Sankey, imbuti e indicatori di riempimento del liquido.

  Considerando un'organizzazione di vendita, una dashboard Looker è in grado di evidenziare gli indicatori che interessano maggiormente come ad esempio il

numero di nuovi utenti acquisiti, l'andamento delle vendite mensili e il numero di ordini da inizio anno. Tali informazioni possono facilitare l'allineamento dei team, identificare le frustrazioni dei clienti e perchè no individuare le perdite di entrate.

Con una dashboard Looker è possibile visualizzare le informazioni anche mediante una serie temporale, per sostenere il monitoraggio dei parametri nel corso del tempo. Looker permette inoltre di tracciare i dati su una mappa per comprendere essi anche da un punto di vista geografico. L'obiettivo di queste funzionalità è aiutare a trarre informazioni utili per prendere decisioni aziendali.

6.1.3 Risultati ottenuti scenario 1

I risultati ottenuti per questo primo scenario riguardano soprattutto la customer base, la quale viene consolidata in quanto non si verifica un elevato Churn Rate ovvero un tasso di abbandono che possa compromettere la sua integrità.

Tramite l'analisi fatta, vengono inoltre applicate attività di retention mirate sia a mantenere i clienti sia ad aumentatare il venduto grazie ad attività di proposition verso gli utenti stessi (tramite ad esempio il volantino personalizzato digitale).

La customer base viene dunque mantenuta attraverso l'incentivo all'acquisto del cliente, sfruttando l'applicazione di promozioni e sconti mirati soprattutto agli utenti dal maggior potenziale di spesa, ossia quelli appartenenti alla classe gold (o alla classe platinum se presente).

6.2 Scenario 2: soluzione fashion basata su Amazon Web Services

6.2.1 Contesto scenario 2

Questo secondo scenario, a differenza del primo, punta ad un cambio di contesto e tecnologia. Nello specifico verrà presentato ed esposto uno use case in contesto fashion basato sulla tecnologia Amazon Web Services. Analogamente a quanto fatto con Google Cloud, di seguito viene illustrata un'immagine dell'architettura

proposta dall'acceleratore, in cui ogni singola fase viene svolta grazie a strumenti focalizzati su componentistiche AWS:

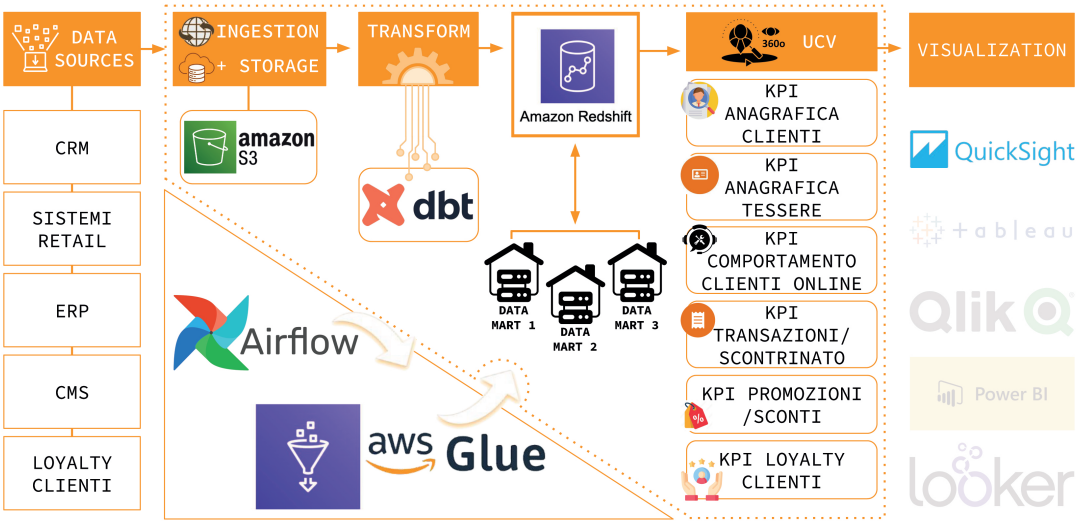


Figura 6.2: Architettura scenario 2
(68)

6.2.2 Obiettivi e strumenti tecnologici scenario 2

In tal caso non si parla più di CRM analitico ma gli obiettivi prefissati (nonostante il contesto differente) non sono poi così differenti dal primo scenario. Infatti, l'idea è sempre quella di mantenere/consolidare la customer base visto che la tesi si focalizza su argomenti incentrati sulla customer analytics. Oltre a ciò, in questa situazione è possibile valutare:

- Analisi e qualità del venduto: attraverso l'analisi delle vendite è possibile ottenere diverse metriche utili all'azienda quali fatturato, prezzo dei prodotti, relative quantità vendute, ricavi netti, costi fissi/variabili e margine lordo/netto delle vendite. Inoltre, un altro modo per portare avanti questo tipo di analisi consiste nel calcolare il venduto su base apertura;
- Performance dei negozi: ciò su cui viene posta l'attenzione riguarda le vetrine da un punto di vista della qualità offerta. Ciò significa che attraverso dei sensori posizionati all'interno di esse, è possibile capire se i potenziali clienti

sono attirati o meno dal punto vendita. Questo lo si fa attraverso il conteggio delle persone che passano davanti alla vetrina, numero che va confrontato con il conteggio delle persone che effettivamente entrano in negozio;

- Considerazione dei fattori esterni: con questo termine si intende ad esempio il meteo oppure la stagionalità. Pensando al meteo, se si considera il periodo estivo e piove spesso durante un certo lasso di tempo, è normale considerare il fatto che ai clienti non vengano venduti costumi. La stagionalità consiste invece in una fluttuazione prevedibile che influenza il comportamento del cliente. Si considerino ad esempio le vacanze di Natale o i periodi di saldi che si ripetono ogni anno. In tali circostanze, la domanda di svariati prodotti incrementerà con ogni probabilità mentre in altri periodi della stagione le vendite caleranno indubbiamente.

Per raggiungere gli obiettivi prefissati, è necessario descrivere ogni strumento impiegato nelle varie fasi dell'architettura AWS (68):

- Apache Airflow: servizio open-source scritto in Python che è usato dai Data Engineers per la creazione, l'organizzazione e il monitoraggio dei flussi di lavoro ma anche per l'esecuzione di pipeline volte all'elaborazione, archiviazione e visualizzazione dei dati.

  Tale soluzione risulta flessibile, scalabile e compatibile con i dati esterni ed è in grado di inviare avvisi/messaggi nel caso in cui l'attività fallisca. Le pipeline sono implementate attraverso l'uso dei grafi aciclici diretti (DAG).

  Airflow è di facile utilizzo poichè l'unico requisito è quello di avere una conoscenza di base del linguaggio di programmazione Python. Inoltre, è uno strumento facilmente integrabile con molti cloud provider.

  In aggiunta, il servizio risulta dinamico, estensibile (adattamento in base all'ambiente aziendale specifico) e scalabile. I motivi che spingono all'utilizzo di Airflow sono sicuramente l'automazione (che ottimizza le performance complessive), la visualizzazione dei processi aziendali, il monitoraggio ed infine il controllo (in tal modo viene permessa l'attuazione di correzioni e la definizione delle responsabilità). (78)

- Glue: servizio di processing dei dati serverless e scalabile (basato su Apache Spark o Python) che supporta tutti i carichi di lavoro in un unico posto centralizzato.

  Attraverso questo strumento è possibile creare, eseguire, monitorare e catalogare pipeline di ETL per il trasferimento dei dati nei data lake. Inoltre, attraverso ulteriori funzionalità è possibile svolgere analisi aggiuntive, fare attività di ML ma anche sviluppare applicazioni.

  A livello di integrazione dati nell'architettura, è tutto più semplificato. Ad esempio, è possibile integrare lo strumento con certi servizi di analisi AWS e con i data lake di S3.

  A prescindere dalle molteplici competenze tecniche che uno può possedere, Glue include interfacce di integrazione e strumenti di creazione di processi semplici e personalizzabili, che sono di facile utilizzo per qualsiasi figura professionale (dagli sviluppatori agli utenti aziendali). (38)

- S3: è un servizio per archiviare oggetti all'interno di bucket dalla durata pressochè illimitata, con funzionalità di replica globale/gestione e con classi di storage convenienti.

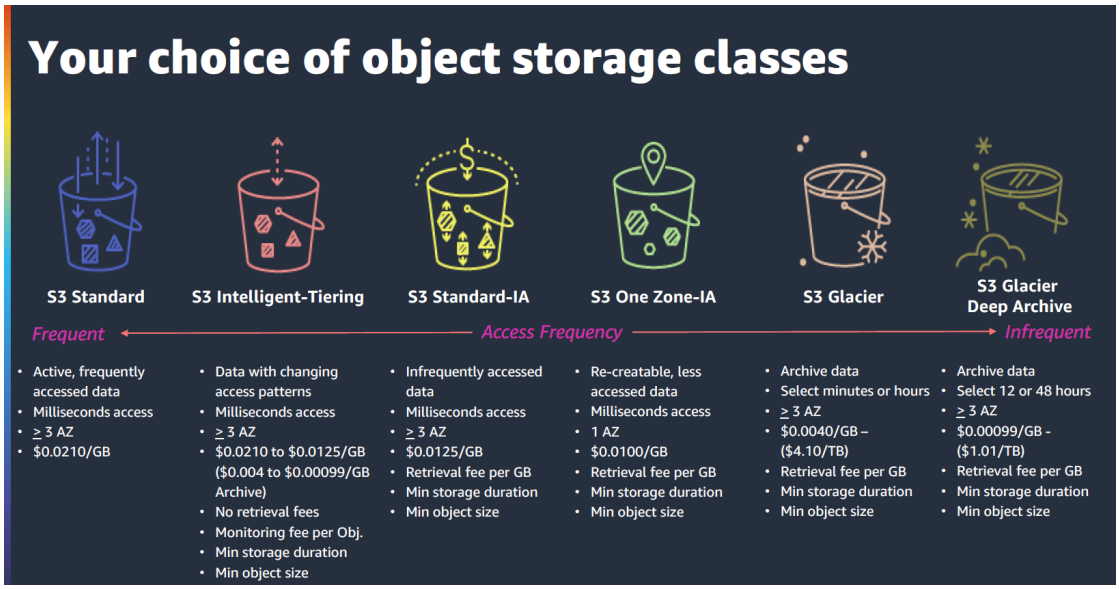


Figura 6.3: S3 AWS
(44)

Un oggetto è l'entità fondamentale di S3 e consiste in un file più tutti i metadati che lo descrivono. Quest'ultimi equivalgono ad un insieme di coppie nome-valore che definiscono l'oggetto. È possibile archiviare un numero illimitato di oggetti in un bucket, che a loro volta non possono però andare oltre una numerosità pari a 100 per ogni account.

Per ogni bucket (container di oggetti/file) creato, bisogna precisare nome e Regione AWS, parametri che in seguito non saranno più modificabili. La scelta di una regione consente l'ottimizzazione della latenza e la riduzione dei costi.

Ad ogni oggetto viene associata sia una chiave (identificatore univoco dell'oggetto nel bucket) sia facoltativamente un ID versione (se il controllo delle versioni S3 è abilitato per il bucket). Dunque si può intendere S3 come una mappa di dati di base tra "bucket + chiave + versione" e l'oggetto stesso.

I bucket/oggetti sono privati e accessibili esclusivamente nel caso in cui vengono concesse in modo esplicito le dovute autorizzazioni d'accesso. In aggiunta, i bucket individuano l'account responsabile del costo di archiviazione e trasferimento dati, e procurano opzioni di controllo degli accessi che possono essere impiegati per la gestione dell'accesso alle risorse di S3. (44)

- Redshift: data warehouse di AWS, non autonomo, che si integra perfettamente con il data lake per mantenere i dati "liberi". Inoltre, esso è completamente gestito con architettura OLAP estremamente parallela e senza condivisione (con archivio dati colonnare scalabile).

  Il ridimensionamento avviene in maniera automatica e le capacità di elaborazione e archiviazione possono scalare in modo indipendente. Inoltre, questo strumento viene protetto con crittografia end-to-end e conformità certificata.

  Redshift è un prodotto altamente performante, scalabile, resiliente, di facile utilizzo e conveniente. Con esso, l'utente può raccogliere informazioni approfondite dai dati in brevissimo tempo, sperimentare prestazioni costantemente elevate e pagare per quello che usa.

  Dunque, ciò di cui si occupa il prodotto è il ridimensionamento automatico, il provisioning del calcolo, il patching automatizzato (alcuni script di codice

aggiornano i sistemi correggendo bug e falle di sicurezza), il failover automatico (processo di trasferimento dei carichi di lavoro ai sistemi di backup), il monitoraggio avanzato, il backup/ripristino, la manutenzione ordinaria e infine la sicurezza/conformità del settore.

- QuickSight: servizio di business intelligence che prevede dei pagamenti esclusivamente per quelle risorse che effettivamente vengono utilizzate (c'è un'ottimizzazione dei costi grazie alle tariffe applicate per ciascuna sessione).

  Come accade per molti servizi presenti in cloud, il ridimensionamento è automatico con alta disponibilità e soprattutto serverless (non è necessario eseguire provisioning in eccesso per i picchi di utilizzo).

  Le descrizioni automatiche che vanno a definire i pannelli di controllo e i report/avvisi possono essere personalizzati e specifici per ogni tipologia di caso d'uso, al fine di ottenere un contesto più dettagliato per ciascun utente.

  L'interfaccia dello strumento è interamente basata sul web e consente di ottenere un'analisi visiva dei dati di facile comprensione, senza richiedere spiccate competenze tecniche di business intelligence.

  Per quanto riguarda il livello di integrazione, c'è nè sia uno a livello di servizi AWS nativi (connettività VPC privata per un accesso AWS sicuro e autorizzazioni IAM native con controllo degli accessi granulare per l'esplorazione dei dati serverless) sia uno a livello di machine learning per l'ottenimento di insight specifici.

  L'accesso ai dati è privato e sicuro ed, inoltre, vengono fatte attività di detection (rilevazione) e forecasting (previsione). La sicurezza è presente sia a livello di riga che di colonna (con supporto API) per il controllo a livello di utente o di gruppo.

  La crittografia viene applicata ai dati end-to-end ma anche ai dati inattivi. Avere una buona sicurezza va senza dubbio ad influenzare positivamente tutto ciò che concerne sia la governance sia tutte le conformità che devono essere integrate. (42)

6.2.3 Risultati ottenuti scenario 2

I risulati ottenuti per questo secondo scenario sono molteplici. Come già accadeva per la GDO con tecnologia Google Cloud, la customer base viene fidelizzata/consolidata attraverso campagne marketing o attraverso la marketing automation (ad esempio, ad un cliente tesserato arriverà qualche promozione).

Quest'ultima fa riferimento a tutte quelle piattaforme software affini al contesto marketing che hanno bisogno di rendere automatizzate attività ripetitive, per poi andarle a sviluppare al fine di renderle più efficaci e controllate. Inoltre, vengono implementate campagne di email marketing e vengono tracciate le attività di clienti attuali e potenziali. (39)

Oltre a porre la giusta e doverosa attenzione verso il cliente, un secondo risultato è mirato all'aumento delle performance dei singoli punti vendita. Attraverso analisi mirate, puntuali e dettagliate è possibile incrementare i rispettivi KPI di negozi che magari all'inizio non sono così performanti. Un paio di conseguenze positive consistono indubbiamente in un netto miglioramento a livello di venduto e perchè no all'acquisizione di nuovi clienti tramite incentivi all'acquisto.

Capitolo 7

Conclusione

Questa tesi ha affrontato l'importante sfida di progettare e sviluppare un acceleratore data analytics (basato su tecnologie cloud) per l'ambito retail, uno dei mercati che può trarre maggiormente vantaggio dalla veloce e profonda evoluzione delle tecnologie e metodologie di una Data Platform, e che inoltre porta con sè tematiche peculiari e molto interessanti della Customer Analytics.

L'acceleratore proposto è stato realizzato in maniera tale per cui l'approccio utilizzato possa essere replicato e standardizzato, al fine di ottimizzare l'efficienza e l'efficacia in termini di implementazione e analisi del dato. In tal caso, l'acceleratore è adattabile sia a diverse tipologie di clienti (contesto GDO, GDS, fashion, beauty) sia a diversi contesti tecnologi (che dipendono dallo specifico cloud provider).

Inoltre, l'acceleratore è una piattaforma perfettamente idonea e appropriata all'assorbimento e alla gestione di tutti i cambiamenti dovuti al progredire dell'evoluzione tecnologica (diffusione dell'IoT, dei Big Data, del cloud, del machine learning, dell'analisi predittiva e dell'AI/GAI).

Dunque, l'obiettivo è quello di non creare debito tecnologico e quindi consentire di stare sempre al passo con i tempi sia a livello di informazioni di dati sia a livello di tecnologia. Coloro che traggono maggior beneficio da questa piattaforma (e che quindi adottano un approccio semplificato e agevolato) sono le figure professionali volte all'analisi del dato, in particolare il data engineer e il data scientist. Tuttavia,

è inevitabile che quest'ultimi si adeguino al cambiamento considerando che la piattaforma non è statica.

In conclusione, questa tesi ha proposto un acceleratore che rappresenta un passo avanti significativo nella standardizzazione delle pratiche analitiche, contribuendo ad una crescita incessante in termini di competitività ed efficienza sul mercato. L'acceleratore avrà sempre una certa flessibilità in risposta sia all'innovazione tecnologica sia alle esigenze mutevoli del panorama retail, garantendo un supporto affidabile e all'avanguardia per le decisioni basate sui dati.

Bibliografia

[1] Batch processing cos'è, . URL https://www.tibco.com/it/reference-center/what-is-batch-processing.

[2] Batch processing vantaggi e casi d'uso, . URL https://www.talend.com/it/resources/batch-processing/.

[3] Batch processing architettura, . URL https://learn.microsoft.com/it-it/azure/architecture/data-guide/big-data/batch-processing.

[4] Big data cosa sono, . URL https://www.oracle.com/it/big-data/what-is-big-data/.

[5] Big data perchè importanti, . URL https://www.sas.com/it_it/insights/big-data/what-is-big-data.html.

[6] Business intelligence software, . URL https://scegliformitore.it/blog/software-di-business-intelligence-i-5-migliori-di-quest-anno\-caratteristiche-e-quale-scegliere/.

[7] Business intelligence perchè è importante, . URL https://www.tableau.com/it-it/learn/articles/business-intelligence.

[8] Business intelligence amazon quicksight, . URL https://aws.amazon.com/it/quicksight/.

[9] Business intelligence google looker, . URL https://injenia.it/google-looker/.

[10] Churn rate come ridurlo, . URL https://www.qualtrics.com/it/experience-management/cliente/churn-rate/.

[11] Churn rate churn analysis, . URL https://www.glossariomarketing.it/significato/churn-rate/.

[12] Cloud perchè usarlo e limitazioni, . URL https://cloud.google.com/learn/advantages-of-cloud-computing?hl=it.

[13] Cloud analytics cos'è, . URL https://www.tibco.com/it/reference-center/what-is-cloud-analytics.

[14] Cloud composer, . URL https://cloud.google.com/composer?hl=it.

[15] Dashboard mockup. URL https://www.visualitics.it/mockup-e-prototipi-nella-creazione-di-dashboard/.

[16] Data lake cos'è, . URL https://cloud.google.com/learn/what-is-a-data-lake?hl=it.

[17] Data processing, . URL https://www.ovhcloud.com/it/learn/what-is-data-processing/.

[18] Data warehouse cos'è, . URL https://www.oracle.com/it/database/what-is-a-data-warehouse/.

[19] Dataproc, . URL https://cloud.google.com/dataproc/docs/concepts/overview?hl=it.

[20] Foto big data, . URL https://www.invisiblefarm.it/i-big-data-enormi-quantita-di-dati/.

[21] Business intelligence, . URL https://www.inside.agency/strumenti-di-business-intelligence/.

[22] Foto cms, . URL https://www.4media.com/article/145,what-is-cms.

[23] Foto crm, . URL https://ascendix.com/blog/who-uses-crm/.

[24] Foto classificazione binaria, . URL https://docs.aws.amazon.com/it_it/machine-learning/latest/dg/binary-classification.html.

[25] Foto confusion matrix, . URL https://blog.csdn.net/flyfish1986/article/details/117741939.

[26] Foto curve di sopravvivenza, . URL https://www.sifact.it/risultati-del-progetto-avvicinare/.

[27] Foto data lake, . URL https://helicaltech.com/services/data-lake-services/.

[28] Foto data warehouse, . URL https://www.chegg.com/homework-help/questions-and-answers/discuss-data-flow-data-warehouse-environment-data-sources-\report-generation-q73200167.

[29] Foto erp, . URL https://www.e-time.it/software-erp-gestionale/.

[30] Intelligenza artificiale, . URL https://www.wired.it/article/intelligenza-artificiale-sensibile-uomo-macchina-lavoro/.

[31] Foto intelligenza artificiale generativa, . URL https://www.techopedia.com/definition/34633/generative-ai.

[32] Foto kernel svm, . URL https://it.mathworks.com/discovery/support-vector-machine.html#:~:text=L'obiettivo%20di%20un%20algoritmo,meno%20nella%20figura%20qui%20sotto.

[33] Foto machine learning, . URL https://humanativaspa.it/machine-learning-come-scegliere-il-modello-migliore/.

[34] Foto rfm, . URL https://www.ecommerce-school.it/blog/scopri-i-tuoi-clienti-migliori-con-la-segmentazione-rfm/.

[35] Foto svm, . URL https://it.mathworks.com/discovery/support-vector-machine.html#:~:text=L'obiettivo%20di%20un%20algoritmo,meno%20nella%20figura%20qui%20sotto.

[36] Foto transazioni scontrinato, . URL https://it.mobiletransaction.org/pos-con-scontrino/.

[37] Foto ttnp, . URL https://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S2224-78902020000300008.

[38] Glue caratteristiche. URL https://docs.aws.amazon.com/it_it/glue/latest/dg/what-is-glue.html.

[39] Marketing automation. URL https://www.digital4.biz/marketing/marketing-automation-cos-e-come-funziona-e-come-scegliere-gli-\strumenti-adatti/.

[40] Modern data platform cos'è. URL https://blog.horsa.com/data-analytics/modern-data-platform-perche-hai-bisogno-di-una-\piattaforma-moderna/.

[41] Personalizzazione statistiche d'ordine. URL https://www.eco.uninsubria.it/VL/VL_IT/sample/sample7.html.

[42] Quicksight cos'è. URL https://aws.amazon.com/it/quicksight/.

[43] Elaborazione real time cos'è e vantaggi. URL https://www.digital-coach.com/it/real-time-analytics/.

[44] S3 come funziona. URL https://docs.aws.amazon.com/it_it/AmazonS3/latest/userguide/Welcome.html.

[45] Ucv crm, . URL https://www.salesforce.com/it/learning-centre/crm/what-is-crm/.

[46] Ucv oltp, . URL https://www.businessintelligencegroup.it/olap-e-oltp-cosa-sono-e-quali-sono-le-principali-differenze/.

[47] Customer lifetime p(active), 2012. URL https://www.slideshare.net/TargetResearch/stima-del-customer-lifetime-value.

[48] Vpn tipologie, 2017. URL https://ostec.blog/en/remote-work/vpn-client-applications/.

[49] Customer lifetime value come misurarlo e perchè è importante, 2019. URL https://www.fluency.cx/risorse/customer-lifetime-value-significato.

[50] Cloud cos'è, 2019. URL https://www.doxee.com/it/blog/tecnologia/i-5-vantaggi-del-cloud-computing/.

[51] Personalizzazione quantili, 2021. URL https://matematicaoltre.altervista.org/statistica-descrittiva-cosa-sono-i-percentili/.

[52] Customer lifetime value costo clienti, 2023. URL https://www.qualtrics.com/it/experience-management/cliente/customer-lifetime-value-clv/?rid=langMatch&prevsite=en&newsite=it&geo=IT&geomatch=.

[53] Customer lifetime perchè è importante, 2023. URL https://www.shopify.com/blog/what-is-customer-lifetime-value.

[54] Churn rate cos'è e come si calcola, 2023. URL https://www.bnova.it/analytics/churn-rate/.

[55] Cloud modelli, 2023. URL https://www.timenterprise.it/approfondimenti/cloud-computing-tipologie-vantaggi.

[56] Data lake architettura, 2023. URL https://www.redhat.com/it/topics/data-storage/what-is-a-data-lake.

[57] Data platform vantaggi e tipologie, 2023. URL https://www.beantech.it/blog/data-platform-cose-quali-vantaggi-offre-e-quante-\tipologie-esistono/.

[58] Git, 2023. URL https://learn.microsoft.com/it-it/devops/develop/git/what-is-git.

[59] Orchestrazione, 2023. URL https://www.databricks.com/it/glossary/orchestration.

[60] Targetizzazione tipologie, 2023. URL https://www.qualtrics.com/it/experience-management/marchio/segmentazione-targeting/.

[61] Marta Abba'. Data platform cos'è, 2021. URL https://www.zerounoweb.it/analytics/data-management/

data-platform-cose-e-perche-garantisce-una-modern-data-\
experience/.

[62] Tania Cerquitelli Elena Baralis. Corso big data business intelligence polito.

[63] Erminia Chiodo. Targetizzazione come individuarla, 2023. URL https://www.ecostampa.it/blog/target-giusto-per-una-campagna-marketing-ecco-la-guida/.

[64] Alessia Delmedico. Personalizzazione collaborative filtering formule, 2019. URL https://tesi.luiss.it/26072/1/701231_DELMEDICO_ALESSIA.pdf.

[65] Thomas Taimre Radislav Vaisman Dirk P. Kroese, Zdravko I. Botev. Data science and machine learning, mathematical and statistical methods. 2020.

[66] Thor Olavsrud John Edwards. Analisi predittiva vantaggi e casi d'uso, 2023. URL https://www.cio.com/article/646967.

[67] Carrie Gray. Customer analytics cos'è. URL https://www.sas.com/it_it/insights/marketing/customer-analytics.html.

[68] Horsa. Materiale azienda horsa.

[69] Barış Karaman. Ttnp ottimizzazione iperparametri, 2019. URL https://towardsdatascience.com/predicting-next-purchase-day-15fae5548027.

[70] Jagreet Kaur. Intelligenza artificiale generativa retail, 2023. URL https://www.xenonstack.com/blog/generative-ai-retail-industry.

[71] Ravindra Kumar. Foto scd2, 2023. URL https://www.linkedin.com/pulse/mastering-slowly-changing-dimensions-scd-data-pyspark-ravindra\-kumar/.

[72] Vito Lavecchia. Cloud vantaggi e svantaggi. URL https://vitolavecchia.altervista.org/principali-vantaggi-e-svantaggi-del-cloud-computing/.

[73] Patrizia Licata. Analisi predittiva cos'è, 2022. URL https://www.digital4.biz/executive/predictive-analytics-cos-e-l-analisi-predittiva-e-come-l-ai-\aiuta-a-prevedere-il-futuro/.

[74] Patrizia Licata. Intelligenza artificiale generativa applicazioni business, 2023. URL https://www.digital4.biz/marketing/generative-ai-che-cosa-e-quali-sono-le-applicazioni-di-\business/.

[75] Jamie Mackie. Personalizzazione puzzle e aree trascurate. URL https://mapp.com/it/blog/dalla-personalizzazione-standard-a-customer-experience-su-\misura/.

[76] Andrea Minini. Test di turing. URL https://www.andreaminini.com/ai/test-di-turing/.

[77] Massimo Montedoro. Business intelligence strumenti, 2021. URL https://universeit.blog/software-business-intelligence/.

[78] Dmitry Nazarevich. Apache airflow caratteristiche, 2022. URL https://innowise.com/it/blog/apache-airflow-introduction/.

[79] Evans Doe Ocansey. Ttnp step da seguire, 2021. URL https://towardsdatascience.com/using-machine-learning-to-predict-customers-next-purchase-day-\7895ad49b4db.

[80] Martina Santoro Luca Papa. Business intelligence cos'è, 2019. URL https://www.digital-coach.com/it/blog/video-blog-marketing/business-intelligence/.

[81] Miriana Piccari. Big data a cosa servono. URL https://its-campus.com/blog/big-data/.

[82] Diana Ramos. Dashboard vantaggi, 2018. URL https://it.smartsheet.com/data-dashboard.

[83] Marco Santandrea. Churn rate quali strategie adottare, 2018. URL https://www.ecommerce-school.it/blog/churn-rate-come-misurarlo-e-quali-strategie-adottare/.

[84] Andrea Serventi. Ucv cos'è, 2021. URL https://blog.mailup.it/2020/05/single-customer-view/.

[85] Elisabetta Severoni. Targetizzazione metriche, 2023. URL https://www.doxee.com/it/blog/customer-experience/una-strategia-personalizzazione-2023/.

[86] Francesco La Trofa. Intelligenza artificiale cos'è. URL https://tech4future.info/intelligenza-artificiale-cose-applicazioni/.

[87] Christy Wilson. Differenza tra batch e real time, 2022. URL https://www.precisely.com/blog/big-data/difference-between-real-time-near-real-time-batch-processing-\big-data.

[88] Arianna Meroni Camilla Zan. Intelligenza artificiale generativa cos'è, 2023. URL https://www.skilla.com/blog/intelligenza-artificiale-generativa-cose-e-alcuni-esempi/.