Università degli Studi di Firenze

Scuola di Ingegneria - Dipartimento di Ingegneria dell'Informazione

Progetto Data Warehouse

Social Media Intelligence: Ottimizzazione del Marketing Attraverso l'Analisi dei Dati

Autrice
Valeria Nardoni

Anno Accademico 2022/2023

Indice

Introduzione

Nell'era digitale in cui viviamo, i social media hanno assunto un ruolo centrale nelle strategie di marketing e comunicazione delle aziende. L'analisi dei dati dei social media è diventata un pilastro fondamentale per comprendere l'interazione tra le aziende e il pubblico, valutare l'efficacia delle campagne di marketing e identificare le tendenze dei consumatori. Nel contesto di questa crescente importanza, il progetto di Data Warehouse mira a esplorare come l'analisi dei dati dei social media può contribuire in modo significativo al successo delle aziende nel marketing online. L'ambiente digitale in cui operano le aziende oggi è caratterizzato da una vasta gamma di piattaforme di social media, ciascuna con il proprio pubblico, stile di comunicazione e metriche di interazione. Le aziende investono notevoli risorse nella gestione delle loro presenze online e nella promozione dei propri prodotti e servizi attraverso le piattaforme social. In questo contesto, l'analisi dei dati dei social media è diventata una pratica essenziale per mantenere un vantaggio competitivo.

Obiettivi

Il progetto si concentra su tre obiettivi principali:

1. **Analisi dell'Engagement del Pubblico**: L'obiettivo è monitorare e analizzare le metriche di engagement del pubblico, come like, con-

divisioni, commenti e clic, per valutare l'interazione con i contenuti pubblicati sulle piattaforme social. Questa analisi aiuterà a comprendere quali tipi di contenuto generano maggiore coinvolgimento da parte del pubblico.

2. **Valutazione dell'Efficacia delle Campagne**: Analizzeremo l'efficacia delle campagne di marketing sui social media, considerando metriche di engagement, conversioni, ROI e il successo nel raggiungere gli obiettivi di marketing. Questa valutazione sarà essenziale per migliorare la strategia di marketing online.

3. **Analisi del Sentimento**: Utilizzeremo analisi del testo e strumenti di analisi del sentimento per valutare il tono e il sentimento dei messaggi sui social media. L'identificazione di sentimenti positivi, negativi o neutri ci consentirà di comprendere come il pubblico percepisce il brand e i suoi messaggi.

Capitolo 1

Analisi dei requisiti

L'analisi dei requisiti è il fondamento su cui costruiremo la solida struttura del nostro Data Warehouse. In questo contesto, i **fatti** rappresentano le interazioni, le misurazioni e le metriche chiave che delineeranno il nostro percorso decisionale, sono le informazioni su cui si vuole ottenere analisi e da cui si desidera estrarre significato. Essi ci permetteranno di catturare l'engagement del nostro pubblico, valutare l'efficacia delle campagne e scoprire le tendenze emergenti. Le **dimensioni di analisi** ci forniranno contesti preziosi, consentendoci di scomporre i dati in categorie significative come piattaforme social media, geolocalizzazione e tipi di contenuto. Forniscono il contesto per i fatti e consentono di rispondere a domande come "chi", "quando", "dove", "come" e "perché".

Questi elementi ci aiuteranno a definire chiaramente le linee guida per progettare, immagazzinare e accedere ai dati in modo ottimale. Nel nostro viaggio di analisi dei dati dei social media e del marketing, l'analisi dei requisiti rappresenta il nostro primo passo verso l'acquisizione di intuizioni profonde e azioni informate.

1.1 Data Mart: Interazioni degli Utenti

1.1.1 Fatti

Costituisce uno dei principali elementi nel contesto in esame. Le interazioni degli utenti comprendono approvazioni, condivisioni, commenti e clic su pubblicazioni e annunci presenti sui social media. Tali manifestazioni di interazione rivestono un'importanza cruciale, in quanto rappresentano l'engagement del pubblico e l'interazione con i contenuti.

1.1.2 Analisi delle dimensioni

Tempo: Una delle dimensioni più critiche in questo contesto è il tempo. La suddivisione del tempo in gerarchie permette di esaminare le tendenze temporali delle interazioni degli utenti. Le gerarchie temporali possono includere:

1. Anno: Per valutare le tendenze annuali.
2. Trimestre: Per analizzare i cambiamenti stagionali.
3. Mese: Per esaminare le variazioni mensili.
4. Giorno: Per una visione giornaliera delle interazioni.
5. Ora: Se necessario, per comprendere l'engagement a livello orario.

Piattaforme Social Media: Questa dimensione permette di confrontare l'engagement su diverse piattaforme di social media. Le piattaforme possono includere Facebook, Twitter, Instagram o altre piattaforme social. Questo aiuta a identificare le differenze nelle interazioni tra piattaforme.

Categorie di Contenuto: La dimensione delle categorie di contenuto è

cruciale per valutare quale tipo di contenuto genera più interazioni. Le categorie possono comprendere notizie, promozioni, eventi o altro ancora.

Geolocalizzazione: Questa dimensione permette di analizzare l'engagement in diverse aree geografiche. È utile per comprendere le differenze regionali nelle interazioni degli utenti.

Segmenti di Utenti: La suddivisione degli utenti in segmenti basati su dati demografici o interessi specifici è importante per confrontare il coinvolgimento tra gruppi. I segmenti possono includere età, genere, interessi e altro.

1.1.3 Misure

Numero di Like: Questa misura quantifica il numero di "Mi piace" ricevuti su pubblicazioni e annunci sui social media. È un indicatore dell'approvazione e dell'apprezzamento dei contenuti da parte del pubblico.

Numero di Condivisioni: Questa misura rappresenta quante volte i contenuti sono stati condivisi da altri utenti. Indica la capacità di un contenuto di raggiungere un pubblico più ampio.

Numero di Commenti: Questa misura conta i commenti ricevuti sulle pubblicazioni. I commenti possono fornire feedback e interazioni dirette con il pubblico.

Numero di Click: Questa misura tiene traccia del numero di click su link all'interno dei contenuti. Rappresenta l'azione diretta degli utenti.

1.1.4 Granularità e Intervallo di Storicizzazione

Nell'ambito delle interazioni sui social media, è possibile considerare una granularità mensile come punto di partenza. Questo comporterebbe la raccolta e l'archiviazione di dati giornalieri relativi alle interazioni degli utenti.

Tale approccio consente di ottenere una visione chiara delle tendenze nel corso del tempo, evitando di sovraccaricare eccessivamente il sistema. L'intervallo di storicizzazione, invece, indica il periodo di tempo durante il quale si desidera conservare i dati storici delle interazioni degli utenti. Questo arco temporale determina quanto indietro nel passato si intendono analizzare le interazioni. La conservazione dei dati storici, nel progetto, è stata considerata per un periodo di cinque anni al fine di esaminare le tendenze a lungo termine.

1.1.5 Carico di lavoro Preliminare

Il carico di lavoro preliminare rappresenta la quantità di dati e di risorse necessarie per l'acquisizione, l'elaborazione e l'archiviazione dei dati iniziali al fine di supportare le analisi delle "Interazioni degli Utenti" sui social media. Questo carico di lavoro comprende le seguenti attività:

Raccolta dei Dati Social Media: Questa attività comprende l'implementazione di strumenti per la raccolta dei dati dai social media. Potrebbe essere necessario stabilire connessioni alle API delle piattaforme social (come Facebook, Twitter, Instagram, etc.) e definire le query per recuperare i dati desiderati, come like, condivisioni, commenti, click e altri tipi di interazioni.

Elaborazione e Trasformazione dei Dati: Una volta raccolti, i dati devono essere elaborati per renderli utilizzabili. Questa fase potrebbe includere la pulizia dei dati per rimuovere informazioni ridondanti o non valide, la normalizzazione dei formati dei dati e la gestione di dati mancanti o incompleti.

Definizione di Politiche di Conservazione dei Dati: Poiché la quantità di dati dei social media può essere significativa, è importante definire politiche di conservazione dei dati che stabiliscano quanto tempo i dati verranno conservati. Questo potrebbe variare in base alle esigenze legali e aziendali.

Monitoraggio e Manutenzione Continua: Una volta che il carico di lavoro iniziale è completato, è importante stabilire procedure di monitoraggio e manutenzione continue per garantire che i dati siano aggiornati e che i sistemi funzionino senza intoppi.

1.2 Data Mart: ROI delle Campagne di Marketing

Il calcolo del rendimento dell'investimento (ROI) delle campagne di marketing riveste una significativa importanza. La valutazione dell'efficacia delle iniziative di marketing online richiede la misurazione del profitto generato in rapporto ai costi sostenuti.

1.2.1 Fatti

Profitto: Questo rappresenta il profitto generato dalle campagne di marketing sui social media. Può includere entrate da vendite, acquisizioni di clienti o altre azioni correlate al marketing.

Costi delle Campagne: Questo rappresenta i costi sostenuti per l'implementazione delle campagne di marketing sui social media. Include spese pubblicitarie, spese di produzione creativa, costi di personale e altri costi associati alle campagne.

ROI (Rendimento dell'Investimento): Questo rappresenta il rapporto tra il profitto generato e i costi delle campagne. Il ROI è una misura chiave per valutare l'efficacia delle iniziative di marketing.

1.2.2 Analisi delle dimensioni

Tempo: Possiamo utilizzare le gerarchie temporali per analizzare il ROI nel tempo, come anni, trimestri, mesi, giorni, ore.

Tipo di Campagna: Questa dimensione ti permette di analizzare il ROI in relazione a diverse tipologie di campagne di marketing, come campagne pubblicitarie, campagne promozionali, campagne di contenuti, ecc.

Piattaforme di Social Media: Puoi esaminare quale piattaforma di social media ha generato il ROI più alto.

Segmenti di Utenti: Questa dimensione ti consente di confrontare il ROI tra diversi segmenti di utenti, come utenti basati su interessi o demografici.

1.2.3 Misure

ROI (%): Questa misura rappresenta il ROI come percentuale, calcolato come:

$$\frac{\text{Profitto} - \text{Costi delle Campagne}}{\text{Costi delle Campagne}} \times 100$$

Questa percentuale indica il ritorno sull'investimento.

Profitto Netto: Questa misura rappresenta il profitto netto generato dalle campagne di marketing, calcolato come: Profitto - Costi delle Campagne

1.2.4 Granularità e Intervallo di Storicizzazione

É stato adottato un approccio basato su una granularità mensile al fine di effettuare l'analisi del ROI su base mensile. Questo approccio consente di analizzare il ROI in relazione a ciascuna campagna specifica, soprattutto se quest'ultima è promossa su diverse piattaforme, consentendo un'analisi dettagliata per ciascuna di esse. Per quanto riguarda l'intervallo di stori-

cizzazione, è importante notare che alcuni dati devono essere conservati per ragioni legali per un periodo di tempo preciso. I dati saranno conservati per il periodo richiesto dalle normative vigenti, e, ove possibile, ad estenderne la durata per scopi di analisi a lungo termine.

1.2.5 Carico di lavoro Preliminare

Raccolta dei Dati delle Campagne: Inizialmente, saranno necessarie risorse per raccogliere i dati relativi alle campagne di marketing sui social media. Ciò potrebbe richiedere l'accesso alle metriche delle campagne dalle diverse piattaforme social media e l'implementazione di strumenti di raccolta dati.

Raccolta dei Dati Finanziari: È necessario raccogliere i dati finanziari relativi alle spese di marketing, che includono le spese pubblicitarie, i costi di produzione creativa, i costi di personale e altre spese correlate alle campagne.

Elaborazione dei Dati: I dati raccolti dalle campagne di marketing e i dati finanziari devono essere elaborati e combinati per calcolare il ROI. Questa attività include la pulizia dei dati, la trasformazione e la normalizzazione dei dati, nonché l'identificazione delle conversioni e dei profitti associati alle campagne.

Archiviazione dei Dati: I dati elaborati devono essere archiviati in modo efficace per supportare le analisi future. Questo potrebbe richiedere l'implementazione di sistemi di archiviazione dati o database appositamente progettati.

Calcolo del ROI: È necessario calcolare il ROI per ciascuna campagna e per l'insieme delle campagne. Questa attività richiede competenze di analisi dei dati e calcolo finanziario.

Documentazione: È importante documentare il processo di raccolta, elaborazione e calcolo del ROI per futuri riferimenti e condivisione con gli stakeholder.

Monitoraggio e Manutenzione Continua: Una volta che il carico di lavoro iniziale è completato, è importante stabilire procedure di monitoraggio e manutenzione continue per garantire che i dati siano aggiornati e che i sistemi funzionino senza intoppi.

1.3 Data Mart: Sentimento dei Messaggi

Nel contesto dell'analisi del sentimenti, va evidenziato che il **sentimento** dei messaggi sui social media può essere considerato un elemento rilevante. Tale parametro rappresenta la percezione relativa al tono dei messaggi, classificandoli come positivi, negativi o neutri.

1.3.1 Fatti

Numero di Messaggi per Sentimento: Questo rappresenta il numero totale di messaggi analizzati, suddivisi in messaggi positivi, negativi e neutri.

1.3.2 Analisi delle dimensioni

Tempo: Utilizza le gerarchie temporali per analizzare come il sentimento dei messaggi cambia nel tempo.

Piattaforme Social Media: Questa dimensione consente di analizzare il sentimento dei messaggi su diverse piattaforme social media, come Facebook, Twitter, Instagram, ecc.

Segmenti di Utenti: Questa dimensione ti consente di confrontare il sentimento dei messaggi tra diversi segmenti di utenti o gruppi demografici.

1.3.3 Misure

Percentuale del Sentimento: Questa misura rappresenta la distribuzione percentuale dei messaggi tra positivi, negativi e neutri.
Sentimento Medio: Questa misura rappresenta la valutazione media del sentimento dei messaggi.

1.3.4 Granularità e Intervallo di Storicizzazione

Consideriamo una granularità mensile e valutiamo, per l'intervallo di storicizzazione, i requisiti di conservazione, lo spazio di archiviazione necessario e cerchiamo di mantenere i dati per un lungo tempo, in modo di poter analizzare le tendenze a lungo termine.

1.3.5 Carico di lavoro Preliminare

Raccolta dei Messaggi: Raccogliere i messaggi dalle diverse piattaforme social media, inclusi i testi dei messaggi e le informazioni correlate come data e ora di pubblicazione.
Classificazione del Sentimento: Utilizzare strumenti di analisi del testo o modelli di Machine Learning per classificare i messaggi come positivi, negativi o neutri in base al loro contenuto.
Elaborazione dei Dati: Pulire e preparare i dati, inclusa la rimozione di duplicati, la normalizzazione dei testi e l'assegnazione delle categorie di sentimento.

Archiviazione dei Dati: Archiviare i dati in modo che siano accessibili per le analisi future.

Calcolo di Misure Chiave: Calcolare misure chiave, come la percentuale di messaggi positivi, negativi o neutri, nonché statistiche sul sentimento medio.

Documentazione: Documentare il processo di raccolta e analisi del sentimento per futuri riferimenti.

Monitoraggio Continuo: Pianificare procedure di monitoraggio continuo per garantire che i dati siano aggiornati e che il sistema di classificazione del sentimento sia affidabile nel tempo.

Capitolo 2

Analisi delle fonti

2.1 Fonti dei dati

Per alimentare i datamart sono stati considerati come fonti il database operazionale dell'azienda e come fonti esterne le API alle piattaformeSocial, dati finaziari della azienda e Integrazione dei dati provenienti da feedback dei clienti e sondaggi aziendali per comprendere il sentiment degli utenti.

2.2 Fonti: Interazione degli Utenti

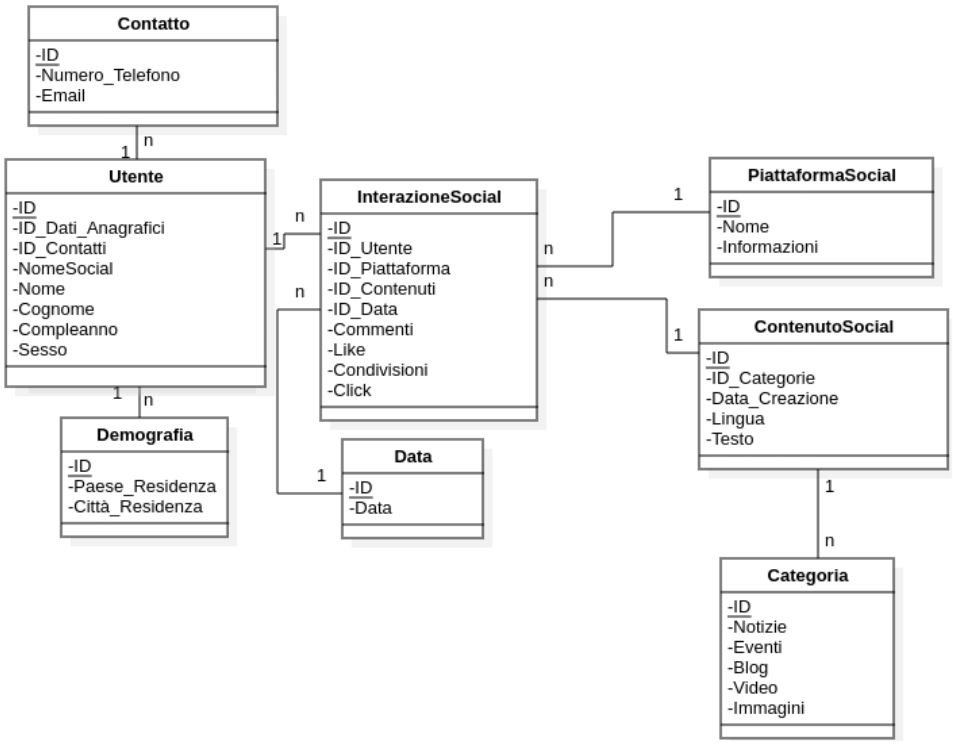

Figura 2.1: Schema del database

2.3 Fonti: ROI delle Campagne di Marketing

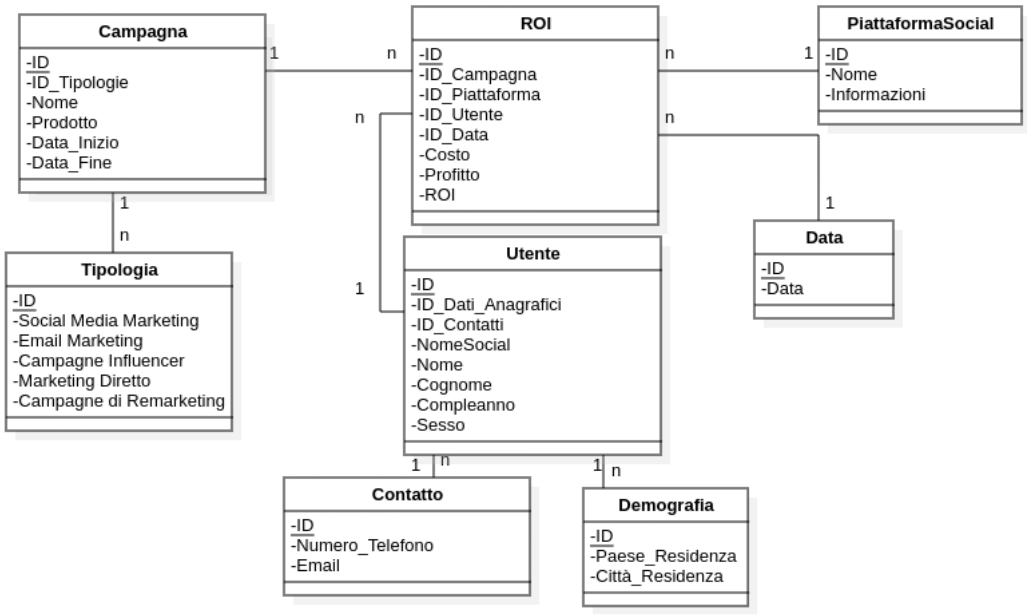

Figura 2.2: Schema del database

2.4 Fonti: Sentimento dei Messaggi

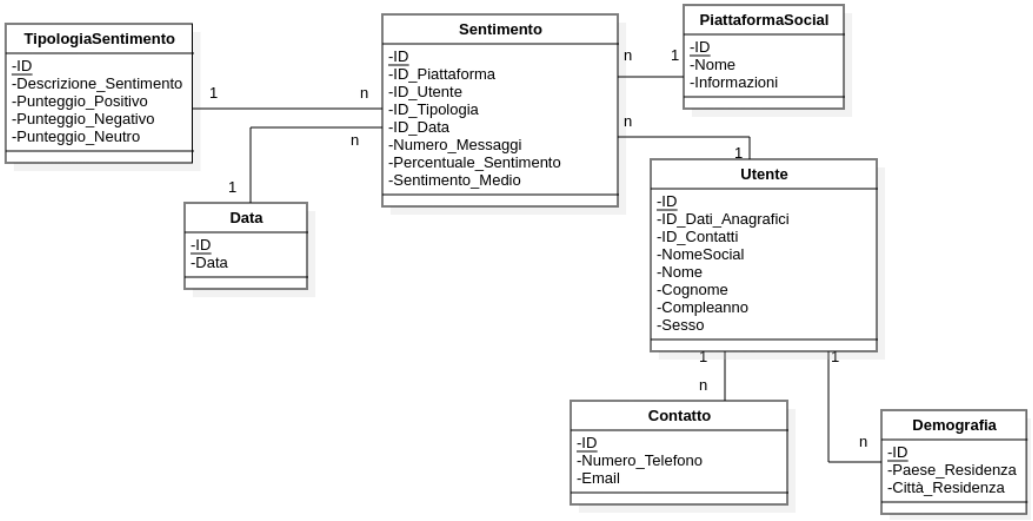

Figura 2.3: Schema del database

2.5 Conciliazione dei dati

La gestione delle possibili discrepanze o conflitti tra i dati provenienti da database diversi rappresenta una fase critica nel processo di progettazione e implementazione di un datamart. Di seguito sono delineate alcune considerazioni per affrontare efficacemente questa problematica:

Standardizzazione dei Dati: Assicurarsi che i dati provenienti da database diversi siano standardizzati secondo regole predefinite. Ad esempio, nelle informazioni demografiche, la codifica dei paesi, delle città e di altri dati simili dovrebbe seguire un modello coerente.

Allineamento delle Chiavi Primarie e Esterne: Garantire che le chiavi primarie e esterne che collegano le tabelle nei vari database siano allineate e

mantenute in modo consistente. Ciò è essenziale per stabilire relazioni significative tra le tabelle all'interno del datamart.

Risolvere Discrepanze di Formato: Trattare eventuali discrepanze nei formati dei dati, come formati di data diversi o unità di misura incoerenti. Applicare trasformazioni durante il processo ETL per uniformare i dati.

Gestione dei Dati Mancanti: Affrontare il problema dei dati mancanti in modo coerente. È possibile optare per l'imputazione dei dati mancanti o decidere di escludere le osservazioni con dati mancanti in base alla situazione specifica.

Documentazione Dettagliata: Documentare dettagliatamente le trasformazioni e le decisioni prese durante il processo ETL. Una documentazione chiara facilita la comprensione delle manipolazioni effettuate sui dati e risolve eventuali discrepanze. Documentare chiaramente la semantica di ogni colonna può aiutare a mitigare possibili problemi di ononimia.

Test e Validazione: Implementare test e procedure di validazione per garantire che i dati nel datamart riflettano con precisione la realtà aziendale. Questo può includere confronti tra i dati nel datamart e le fonti originali.

Aggiornamento Periodico: Rivedere e aggiornare regolarmente il processo ETL e le regole di gestione dei dati per riflettere cambiamenti nelle fonti di dati o nelle esigenze aziendali.

Affrontare pro-attivamente questi aspetti durante la progettazione del datamart contribuirà a minimizzare i conflitti e a garantire che il datamart sia una risorsa affidabile per l'analisi aziendale.

Capitolo 3

Progettazione concettuale

Per la progettazione concettuale dei tre datamart, è utile considerare diversi aspetti, inclusi i concetti di Fatto, Misura, Dimensione, e l'eventuale implementazione di gerarchie.

Fatto: Il fatto rappresenta un evento o una transazione che si verifica in un'organizzazione e che si desidera analizzare. In un datamart, la tabella dei fatti contiene le misurazioni numeriche o quantitative correlate a questo evento. Un fatto esprime una associazione molti-a-molti tra le dimensioni.

Misura: Una misura è una quantità numerica che rappresenta il valore o l'importanza di un dato. Le misure sono tipicamente collegate agli eventi registrati nella tabella dei fatti.

Dimensione: Le dimensioni sono gli aspetti con cui si desidera analizzare i dati nella tabella dei fatti. Sono attributi che forniscono contesto alle misurazioni. Le tabelle dimensionali contengono i dettagli su queste dimensioni.

Gerarchie: Le gerarchie rappresentano relazioni ordinate tra i dati. In un contesto dimensionale, le gerarchie sono spesso strutturate come livelli di dettaglio e aggregazione.

3.1 DFM: Interazione degli Utenti

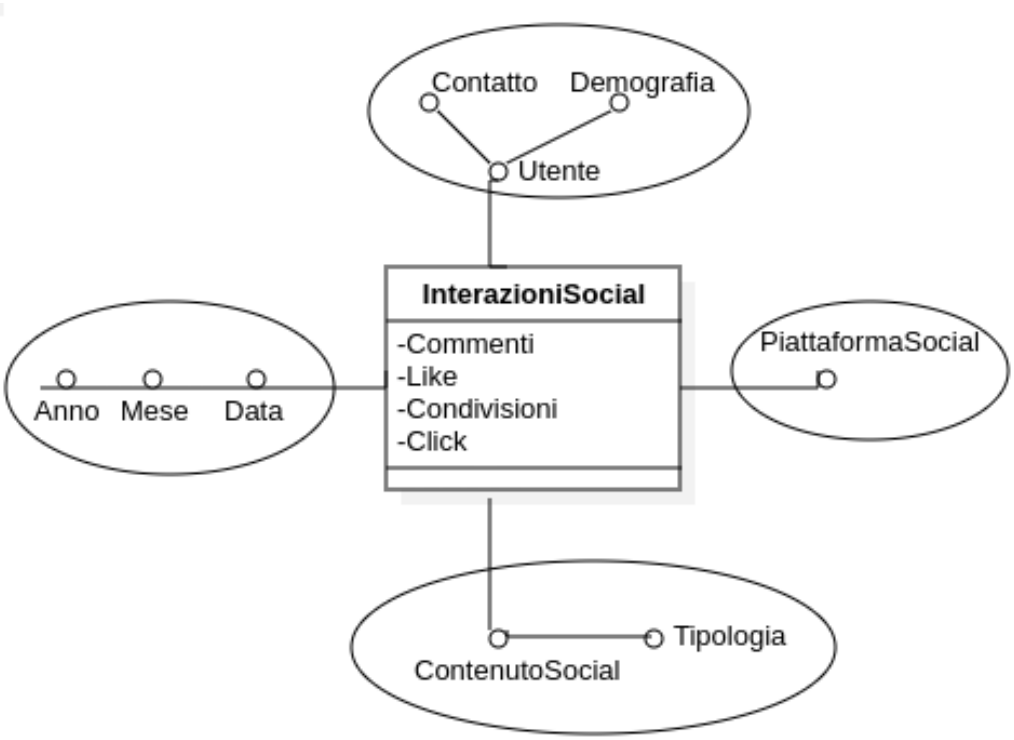

Figura 3.1: DFM: Interazione degli Utenti

```
Le misure di interesse sono:
Commenti: unitario
Like: unitario
Condivisioni: unitario
Click: unitario
```

Queste metriche sono considerate unitarie perché misurano la singola azione o interazione dellútente con il contenuto sociale. Considero anche misure aggregate per conoscere i numeri totali dell'iterazioniSocial.

Il presente DFM è stato progettato per rappresentare in modo efficace le iterazioni sociali, enfatizzando le misure chiave e le dimensioni pertinenti.

Il fatto principale, **InterazioneSociali**, si concentra sulle attività coinvolte nelle interazioni sociali, quali commenti, like, condivisioni e click. Queste misure forniscono insight preziosi sul coinvolgimento degli utenti con il contenuto. Le dimensioni selezionate sono state scelte per catturare gli aspetti temporali, utente, contenuto sociale e piattaforma sociale. La gerarchia temporale offre la possibilità di analizzare le interazioni sociali nel contesto del tempo, fornendo una visione cronologica delle attività. La dimensione utente è suddivisa ulteriormente in demografia e contatti, consentendo un'analisi dettagliata delle caratteristiche degli utenti e delle informazioni di contatto. La gerarchia della dimensione "ContenutoSocial", che include la categoria del contenuto, aggiunge un livello di dettaglio significativo. Questo permette di esplorare le preferenze degli utenti in relazione ai diversi tipi di contenuto sociale. Allo stesso modo, la dimensione "PiattaformaSocial" consente di distinguere le interazioni sociali tra diverse piattaforme. Complessivamente, questo DFM offre un quadro completo delle interazioni sociali, integrando misure significative con dimensioni che consentono analisi approfondite. La struttura gerarchica facilita la navigazione e l'interpretazione dei dati, fornendo una base solida per l'analisi e l'ottimizzazione delle strategie di coinvolgimento sui social media. In sintesi, il Datamart delle Interazioni Sociali è progettato per consentire una comprensione dettagliata del comportamento degli utenti sulle piattaforme sociali, permettendo analisi approfondite su diversi aspetti delle interazioni sociali.

3.2 DFM: ROI delle Campagne di Marketing

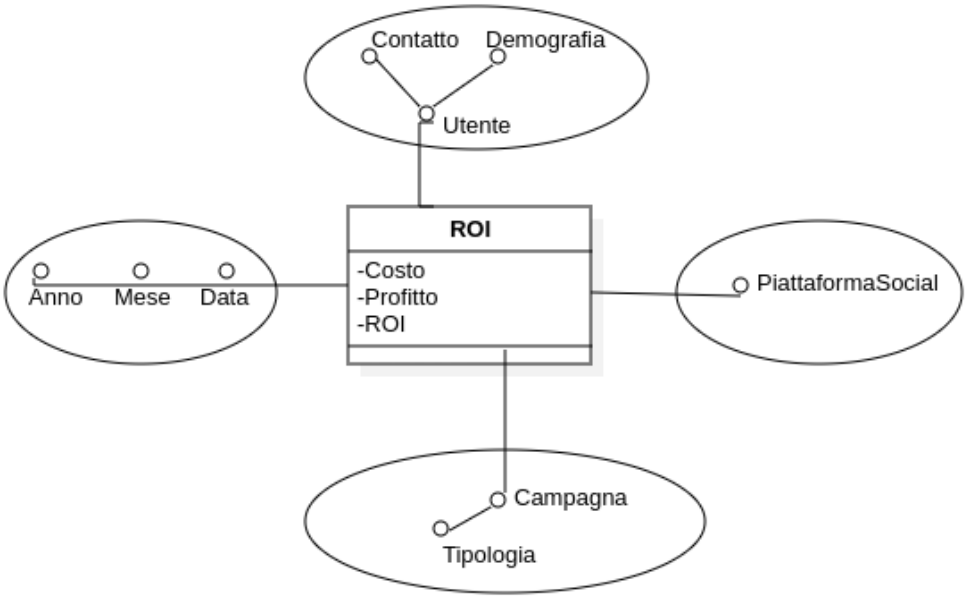

Figura 3.2: DFM: ROI delle Campagne di Marketing

```
Le misure di interesse sono:
Costo: Dettaglio
Profitto: Derivato
ROI: Derivato
```

Il costo è un attributo di dettaglio, ogni riga di dati ha un valore specifico. Profitto e ROI sono derivati, il primo considera anche l'acquisizione di nuovi clienti.

Il presente DFM è stato progettato per fornire un'analisi approfondita delle performance delle campagne di marketing, concentrandosi sulle metriche chiave legate al ritorno sugli investimenti (ROI). Il fatto principale, **ROI**, è centrale nell'analisi finanziaria delle campagne e include misure fondamentali

come costo, profitto e ROI percentuale. Le dimensioni selezionate forniscono un contesto ricco per interpretare il ROI in modo dettagliato. La gerarchia temporale, "Data", consente di esplorare le variazioni delle performance nel tempo, identificando trend e pattern significativi. La dimensione "Utente" è suddivisa ulteriormente in demografia e contatto, offrendo una prospettiva dettagliata sulle caratteristiche degli utenti coinvolti nelle campagne. La gerarchia "Campagna" cattura la tipologia delle campagne di marketing, consentendo di valutare l'efficacia delle diverse strategie di marketing. Questa dimensione è legata alla gerarchia "PiattaformaSocial", che distingue le performance delle campagne su diverse piattaforme sociali. In termini pratici, questo DFM consente di rispondere a domande chiave come "Quali campagne hanno generato il miglior ROI?" o "Come varia il ROI in base alle piattaforme sociali utilizzate?". La struttura gerarchica agevola l'analisi di fattori chiave che influenzano il successo delle campagne di marketing. In sintesi, il DFM per il Datamart ROI delle Campagne di Marketing offre una prospettiva completa sull'efficacia finanziaria delle campagne, integrando misure cruciali con dimensioni significative per supportare decisioni informate e ottimizzazione delle strategie di marketing.

3.3 DFM: Sentimento dei Messaggi

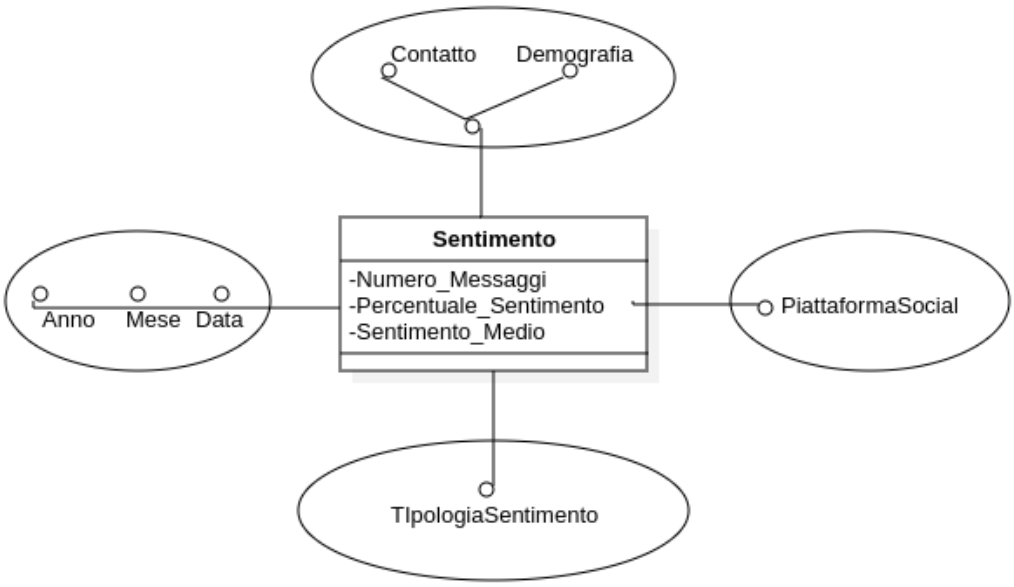

Figura 3.3: DFM: Sentimento dei Messaggi

```
Le misure di interesse sono:
Numero_Messaggi: Dettaglio
Percentuale_Sentimento: Numerico
Sentimento_Medio: Numerico
```

Il presente DFM è stato concepito per analizzare approfonditamente il sentimento associato ai messaggi presenti sulle piattaforme sociali, fornendo metriche chiave per misurare l'atteggiamento degli utenti. Il fatto principale, **Sentimento**, è centrale nell'analisi del mood generale, e le misure incluse, come "Numero Messaggi", "Percentuale di Sentimento" e "Sentimento Medio", offrono una panoramica completa sull'andamento emotivo nel tempo. Le dimensioni selezionate consentono di contestualizzare il sentimento e di

esplorare dettagliatamente le influenze che possono contribuire alle variazioni nei dati. La gerarchia temporale, "Data", è fondamentale per rilevare trend nel sentimento nel corso del tempo, identificando eventi o periodi significativi. La dimensione "TipologiaSentimento" cattura le diverse categorie o interpretazioni del sentimento, arricchendo l'analisi con una comprensione più fine delle sfumature emotive. Questo collegamento permette di esaminare il modo in cui le diverse tipologie di sentimenti contribuiscono alla composizione complessiva del mood. La gerarchia "Utente" si suddivide ulteriormente in "Contatto" e "Demografia", consentendo di analizzare come le caratteristiche degli utenti e i loro contatti possono influire sulle espressioni emotive. La connessione con "PiattaformaSocial" offre un'ulteriore dimensione per comprendere come il sentimente varia tra diverse piattaforme. Questo DFM è progettato per rispondere a domande chiave come "Quali tipologie di sentimenti sono più prevalenti?", "Come varia il sentimento in base alla demografia degli utenti?" e "Esiste una correlazione tra il sentimento e le diverse piattaforme sociali?". In sintesi, il Datamart del Sentimento offre una visione completa e dettagliata sull'andamento emotivo, permettendo analisi approfondite e informate sull'atteggiamento degli utenti sulle piattaforme sociali.

Capitolo 4

Progettazione logica

In presenza di un database con dati di complessità variabile nel tempo viene richiesta flessibilità, quindi il modello logico scelto è quello del **ROLAP** (Relational OLAP):

- **Struttura basata su database relazionali:** ROLAP si basa su database relazionali standard.
- **Flessibilità e scalabilità:** ROLAP è noto per la sua flessibilità nel gestire grandi volumi di dati e per adattarsi a cambiamenti nella struttura dei dati. È particolarmente adatto per ambienti in cui la struttura dei dati è complessa o variabile nel tempo.
- **Performance:** La performance in ROLAP può variare a seconda dell'implementazione e della complessità delle query. Un'adeguata ottimizzazione delle query è essenziale per garantire prestazioni accettabili.

Il modello viene riadattato usando lo **Star Schema**.

4.1 Star Schema

L'obiettivo della fase di progettazione logica è la generazione di uno schema logico che minimizzi il tempo di risposta delle query presenti nel workload nel rispetto dei vincoli di spazio e di tempo. Il passo principale della progettazione logica è la traduzione degli schemi di fatto in schemi logici, come per esempio lo Star schema.
L'obiettivo principale dello star schema è facilitare le query analitiche e semplificare l'accesso ai dati per l'analisi. Nello star schema, si distinguono due tipi principali di tabelle:

- **Tabella Fatto (Fact Table):** Rappresenta i fatti o le misure principali dell'analisi. Ogni riga nella tabella fatto rappresenta un'istanza di misura.

- **Tabelle Dimensionali (Dimension Tables):** Rappresentano le dimensioni associate ai fatti. Ogni riga in una tabella dimensionale rappresenta un'istanza di una dimensione.

Ogni DT è caratterizzato da una chiave primaria (chiave surrogata) e da un insieme di attributi che descrivono le dimensioni di analisi a diversi livelli di aggregazione. La chiave della FT è composta dalle chiavi delle varie DT, inoltre contiene un attributo per ogni misura. L'accesso ai dati e la visione multidimensionale dei dati avviene tramite join tra la FT e DT.

4.2 Star Schema: Interazioni degli Utenti

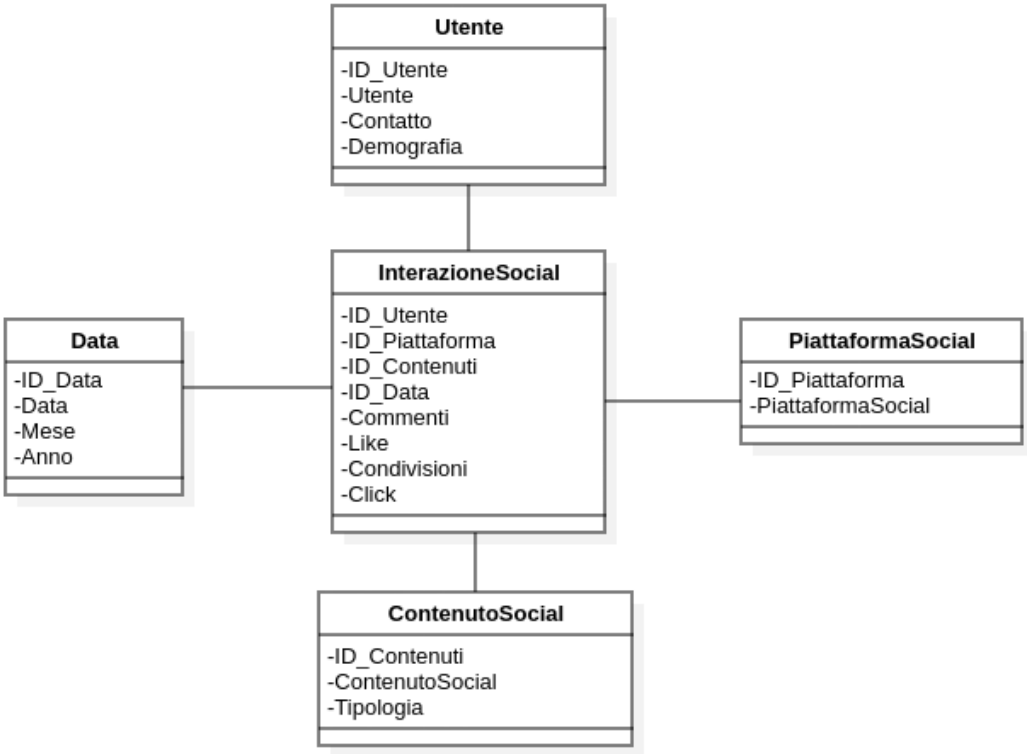

Figura 4.1: Star schema: Interazione degli utenti

4.3 Star schema: ROI delle Campagne di Marketing

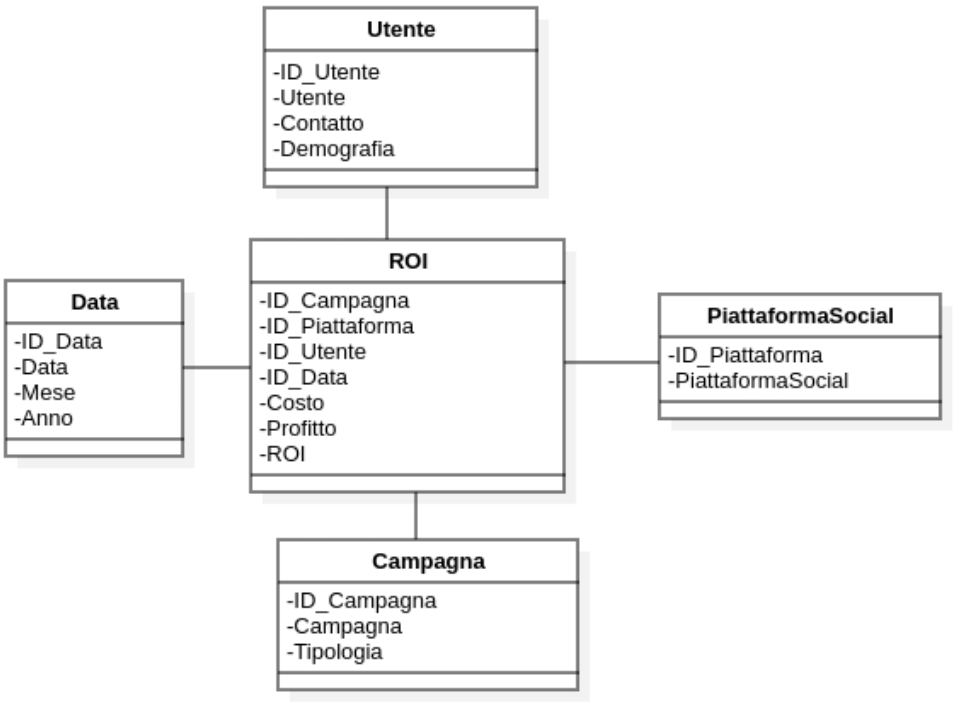

Figura 4.2: Star schema: ROI delle Campagne di Marketing

4.4 Star schema: Sentimento dei Messaggi

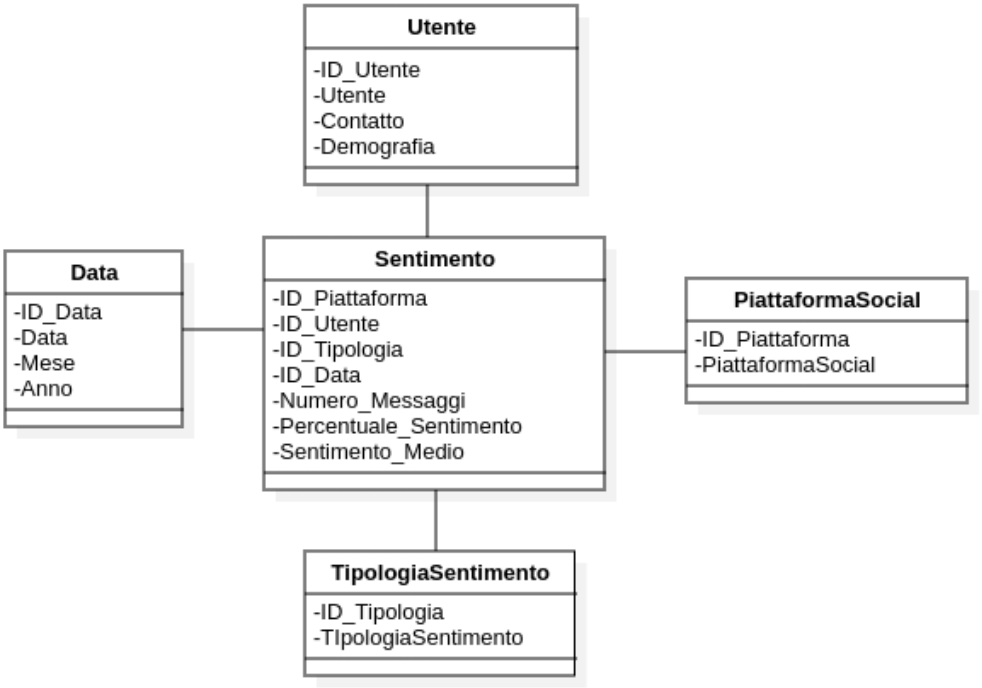

Figura 4.3: Star schema: Sentimento dei Messaggi

Capitolo 5

Progettazione fisica

Nella fase di progettazione fisica di un database, l'attenzione si focalizza su come implementare concretamente lo schema logico progettato nella fase precedente. Uno degli aspetti che può essere preso in considerazione sono le **viste**. Le viste in un database rappresentano un modo di visualizzare i dati da una o più tabelle in modo simile a una tabella virtuale. Sono salvate come oggetti nel database, ma i dati effettivi non vengono duplicati.Nel complesso, la decisione di utilizzare viste dipende dalle esigenze specifiche del sistema e dalle richieste degli utenti. Le viste possono essere uno strumento utile per semplificare l'accesso ai dati, migliorare le prestazioni delle query e facilitare la manutenzione, ma è importante bilanciare questi benefici con le considerazioni di costo e complessità.

Viste Primarie: Le viste primarie corrispondono direttamente alla tabella dei fatti principale nel modello di data warehouse. Questa tabella rappresenta il modello più fine, senza aggregazioni significative. Le viste primarie forniscono una rappresentazione dettagliata dei dati di base senza alcuna aggregazione. Sono particolarmente utili quando gli utenti necessitano di analizzare dati al livello più dettagliato possibile.

Viste Secondarie: Le viste secondarie sono correlate a modelli o pattern secondari che coinvolgono aggregazioni dei dati. Queste viste vengono create a partire dalle viste primarie o, in alcuni casi, direttamente dai dati operazionali. Le viste secondarie contengono dati aggregati, sintetizzati o trasformati che rispondono a domande analitiche specifiche. Sono progettate per semplificare l'analisi delle informazioni a livelli superiori di aggregazione e fornire una visione più sintetica e strategica.

Il problema della materializzazione delle viste viene pertanto formulato come:
Identificare l'insieme delle viste che minimizzano il tempo di risposta al carico di lavoro nel rispetto dei vincoli di sistema (spazio su disco e tempo di aggiornamento).
Nel contesto della progettazione del data warehouse, è stata scelta un'architettura centralizzata che utilizza un singolo database per gestire il data warehouse e le viste associate. Questa decisione è stata guidata da diversi fattori, tra cui la centralizzazione dei dati, la semplificazione della progettazione e la facilità di accesso ai dati per analisi.Le viste sono state implementate all'interno del database del data warehouse per semplificare l'accesso ai dati e migliorare le prestazioni delle query. Queste viste includono sia viste primarie che forniscono dettagli atomici, sia viste secondarie che coinvolgono aggregazioni strategiche per supportare analisi a livello superiore.

5.1 Viste: Interazioni degli Utenti

Vista Primaria: La vista principale è un'istantanea completa delle interazioni sociali. Include informazioni come chi ha interagito, su quale piatta-

forma, con quale contenuto e le metriche specifiche di engagement come like, condivisioni, commenti e click.La struttura della vista principale deve essere progettata in modo da coprire tutte le dimensioni e le misure necessarie per l'analisi. Ad esempio, potrebbe coinvolgere le tabelle Utente, Piattaforma-Social, IterazioneSocial e Data

Mantenendo la vista principale il più informativa possibile e sfruttando le operazioni di aggregazione e manipolazione, direttamente su di essa, semplifichiamo il modello dati e miglioriamo le prestazioni.Tuttavia, in alcune situazioni, potrebbe essere utile creare viste secondarie per semplificare e migliorare la comprensione delle interrogazioni.
Viste Secondarie:

- **Totale di Engagement per Utente:**Una vista che aggrega il totale delle metriche di engagement per ciascun utente. La cardinalità di questa aggregazione sarebbe "uno a molti", poiché un utente può avere molte interazioni sociali.

- **Totale di Engagement per Piattaforma:**Una vista che aggrega il totale delle metriche di engagement per ciascuna piattaforma. La cardinalità di questa aggregazione sarebbe anch'essa "uno a molti", poiché una piattaforma può avere molte interazioni sociali.

5.2 Viste: ROI delle Campagne di Marketing

Vista Primaria: La vista principale potrebbe includere dettagli sulle campagne di marketing, le interazioni sociali associate a tali campagne, informazioni sugli utenti e sulle piattaforme coinvolte. Coinvolge le tabelle

ROI, Campagna, Utente, PiattaformaSocial

Vista Secondaria:

- **ROI totale per ciascuna campagna**: Consentirebbe di analizzare il rendimento delle campagne a un livello più alto di astrazione. Cardinalità "uno a uno" , ogni riga della vista corrisponde a una specifica campagna.

5.3 Viste: Sentimento dei Messaggi

Vista Primaria: La vista principale incorpora tutte le informazioni chiave necessarie per analizzare il sentimento dei messaggi, inclusi dati temporali, utenti, contenuti, piattaforme e metriche relative al sentimento. Coinvolge le tabelle Sentimento, PiattaformaSocial, Data, Utente

Vista Secondaria:

- **Percentuale di Sentimento per Piattaforma**: Vista che aggrega la percentuale di sentimento per ogni piattaforma. Cardinalità "uno a molti", poiché una piattaforma può avere più messaggi.
- **Analisi del Sentimento per Data**: Vista che analizza il sentimento al variare della data. Cardinalità "uno a molti", poiché per ogni data ci possono essere molteplici espressioni o indicazioni di sentimenti diversi.

5.4 Indici

Gli indici in un database sono strutture che migliorano la velocità di ricerca e recupero dei dati. Facilitano l'accesso ai dati in base a determinati criteri, accelerando le query e migliorando le prestazioni complessive del sistema. Alcune esigenze da considerare:

- Costruire indici su attributi non chiave delle DT per accelerare le operazioni di selezione
- Costruire indici sulle chiavi esterne delle FT per accelerare l'esecuzione delle join
- Costruire indici sulle misure delle FT

5.5 Indici: Interazioni degli Utenti

Nella FT **InterazioniSocial** è presente la chiave primaria composta 4 attributi, che rappresentano le chiavi surrogate delle 4 dimensioni. La chiave primaria permette di ottimizzare le join sulle tabelle delle dimensioni. In base alla scelta delle viste secondarie, considererei la possibilità di indicizzare gli attributi PiattaformaSocial, Utente delle rispettive tabelle.

5.6 Indici: ROI delle Campagne di Marketing

Nella FT **ROI** è presente la chiave primaria composta 4 attributi, che rappresentano le chiavi surrogate delle 4 dimensioni. La chiave primaria permette di ottimizzare le join sulle tabelle delle dimensioni. Considererei un indice sull'attributo Campagna sulla sua rispettiva DT e sull'attributo Utente.

5.7 Indici: Sentimento dei Messaggi

Nella FT **Sentimento** è presente la chiave primaria composta 4 attributi, che rappresentano le chiavi surrogate delle 4 dimensioni. La chiave primaria

permette di ottimizzare le join sulle tabelle delle dimensioni. Considererei un indice sull'attributo PiattaformaSocial e Data.

Capitolo 6

Progettazione alimentazione

La progettazione dell'alimentazione (o processo di ETL - Extract, Transform, Load) in un data warehouse coinvolge la fase di estrazione dei dati da diverse fonti, la loro trasformazione secondo le esigenze del sistema e il caricamento nei data mart o nel data warehouse stesso. In questa fase, si possono implementare due approcci principali per estrarre i dati dalle fonti: l'estrazione statica e l'estrazione incrementale.

La scelta tra estrazione statica e incrementale dipende dalle esigenze specifiche del sistema e dal volume dei dati. Se il volume dei dati è relativamente piccolo e le modifiche sono rare, l'estrazione statica potrebbe essere sufficiente. In caso contrario, con grandi volumi di dati e l'esigenza d'ottimizzare le risorse è opportuna l'estrazione incrementale. Nel caso del progetto lavoriamo con una grande quantità di dati dinamici nel tempo.

L'estrazione incrementale è una strategia di estrazione dei dati che prevede l'identificazione e l'estrazione solo dei dati che sono stati modificati o aggiunti dalla precedente estrazione. Questo approccio è utile per ridurre il carico sulle risorse e accelerare il processo di ETL.

6.1 Estrazione Incrementale: Basata su marche temporali

L'implementazione basata su marche temporali è più semplice da comprendere e da gestire rispetto all'implementazione di trigger. Le marche temporali forniscono un approccio chiaro e intuitivo per tenere traccia delle modifiche nei dati. L'implementazione basata su marche temporali è più semplice da comprendere e da gestire rispetto all'implementazione di trigger. Le marche temporali forniscono un approccio chiaro e intuitivo per tenere traccia delle modifiche nei dati. La gestione delle marche temporali è generalmente più semplice e richiede meno manutenzione rispetto a configurare e gestire trigger complessi. Riducendo la complessità, si semplifica anche la manutenzione del sistema nel tempo

Gli svantaggi da tenere conto sono un certo periodo di latenza nell'identificazione delle modifiche, a seconda di quando vengono effettuate le marche temporali e la gestione accurata delle marche temporali richiede l'attenzione per garantire che vengano correttamente mantenute. Nei tre casi dei Data Mart considerati la prima estrazione è statica, successivamente saranno basate su Marche Temporali.

6.2 Estrazione: Interazione degli Utenti

```
-- Prima Estrazione
SELECT *
FROM InterazioniSociali;
```

```
-- Estrazione Marche Temporali
SELECT *
FROM InterazioniSociali
WHERE DataModifica > 'ultima_data_estratta';
```

Nel codice sopra, DataModifica rappresenta la marca temporale che tiene traccia dell'ultima modifica apportata a ciascun record. Durante l'estrazione incrementale, vengono selezionati solo i record che sono stati modificati dopo l'ultima estrazione.

Il sistema prevede un processo di aggiornamento delle marche temporali notturno, eseguito giornalmente durante le ore notturne. Durante l'aggiornamento notturno, il sistema identifica e registra tutte le modifiche apportate ai dati dal giorno precedente. Le marche temporali vengono aggiornate di conseguenza per riflettere gli ultimi cambiamenti. le considerazioni effettuate su questo Data Mart sono valide anche per i successivi Data Marts.

6.3 Estrazione: ROI delle Campagne di Marketing

```
-- Prima Estrazione
SELECT *
FROM ROI_CampagneMarketing;
```

```
-- Estrazione Marche Temporali
SELECT *
FROM ROI_CampagneMarketing
WHERE DataModifica > 'ultima_data_estratta';
```

6.4 Estrazione: Sentimento dei Messaggi

```
-- Prima Estrazione
SELECT *
FROM SentimentoMessaggi;
```

```
-- Estrazione Marche Temporali
SELECT *
FROM SentimentoMessaggi
WHERE DataModifica > 'ultima_data_estratta';
```

Capitolo 7

Visualizzazione dati

Attraverso l'implementazione di Django, ho sviluppato un codice efficace per generare e gestire un database relazionale. La popolazione dei dati è stata eseguita mediante la generazione casuale, offrendo una base eterogenea per le analisi successive. L'impiego della libreria pandas ha agevolato la manipolazione e l'analisi dei dati, mentre l'utilizzo di matplotlib mi ha fornito gli strumenti per creare visualizzazioni grafiche esplicative.
L'approccio di generazione casuale dei dati mi ha consentito di esplorare vari casi d'interesse, offrendo un'ampia prospettiva dei possibili scenari. L'analisi e l'interpretazione dei dati sono facilitate grazie alla potenza delle librerie adottate.

7.1 Grafici: Interazione degli Utenti

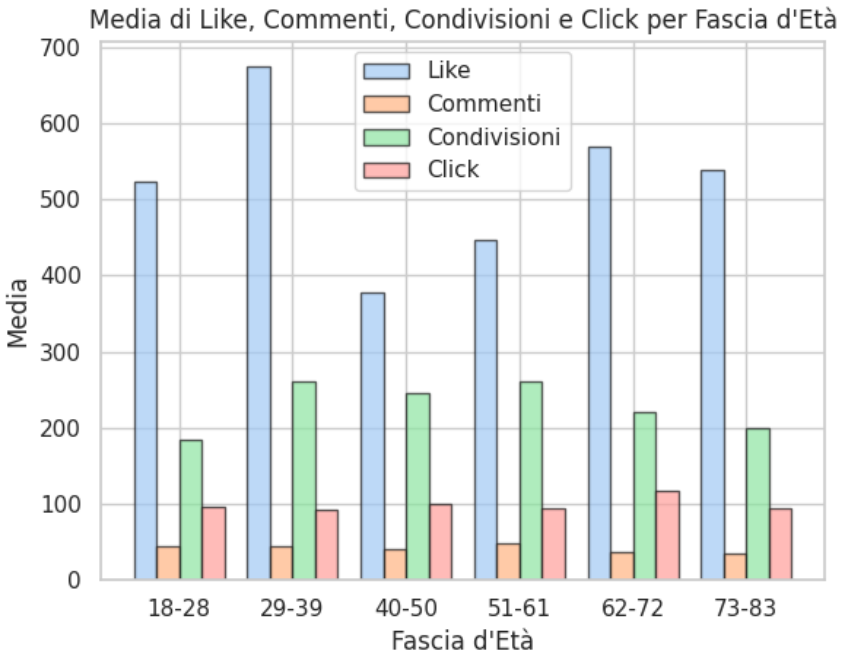

Figura 7.1: Media delle Misure d'interesse

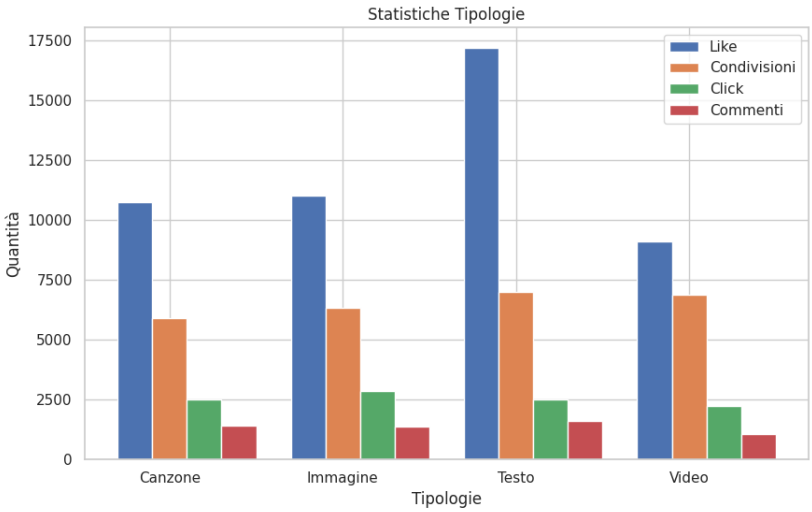

Figura 7.2: Somma delle Misure d'interesse al variare delle Tipologie

Nelle Fig.7.1 e Fig.7.2, stiamo attualmente analizzando i dati considerando l'intero anno 2023 come intervallo temporale di riferimento. Tuttavia, è importante notare che nella Fig.7.3, le informazioni sono state estratte esclusivamente dai dati ricevuti nei mesi di novembre e dicembre. Questo approccio temporale più ristretto fornisce uno sguardo più dettagliato e specifico sulle dinamiche dei dati durante questo periodo specifico, consentendo un'analisi più focalizzata e approfondita per questi due mesi in particolare.

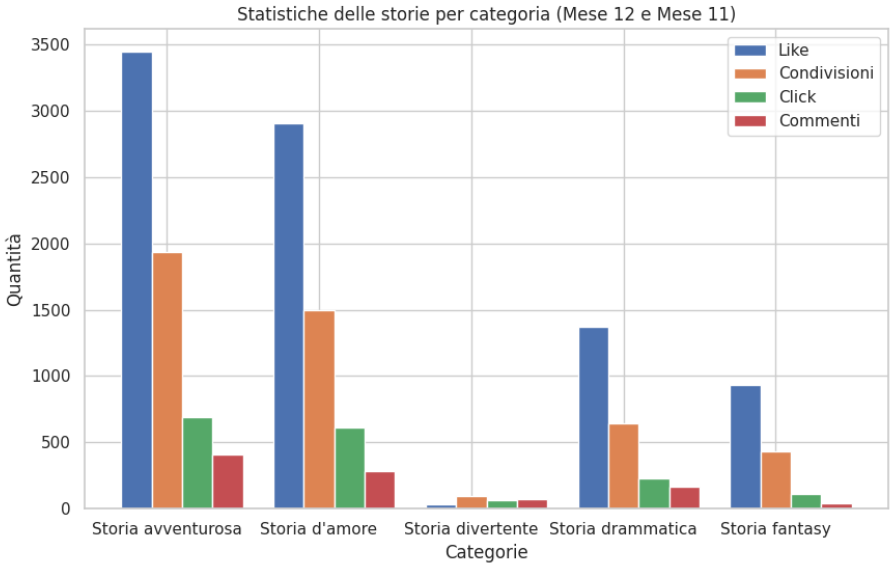

Figura 7.3: Somma delle Misure d'interesse al variare delle Categorie, nei mesi di novembre e dicembre

Esempio Query:

```
df = pd.read_csv('risultato_query.csv')

query = df[(df['click'] < 20)]

risultato_finale = query[['categoria', 'tipologia', 'nome_x',
    'cognome', 'eta']]

print(risultato_finale)
```

	categoria	tipologia	nome_x	cognome	età
1	Storia drammatica	Testo	Tammy	Keith	35
8	Storia fantasy	Video	Sheila	Cooper	52
20	Storia fantasy	Canzone	Heather	Nichols	33
25	Storia divertente	Testo	April	Nguyen	34
29	Storia d'amore	Canzone	Jeffrey	Lawrence	67
38	Storia drammatica	Testo	Debra	Christensen	33
61	Storia drammatica	Immagine	Evan	Jimenez	78

Figura 7.4: Studio su post con pochi click

7.2 Grafici: ROI delle Campagne di Marketing

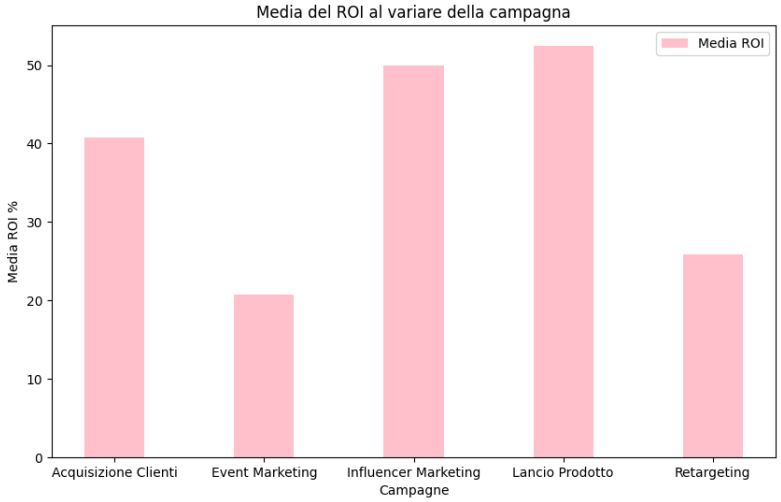

Figura 7.5: Media ROI al variare delle campagne

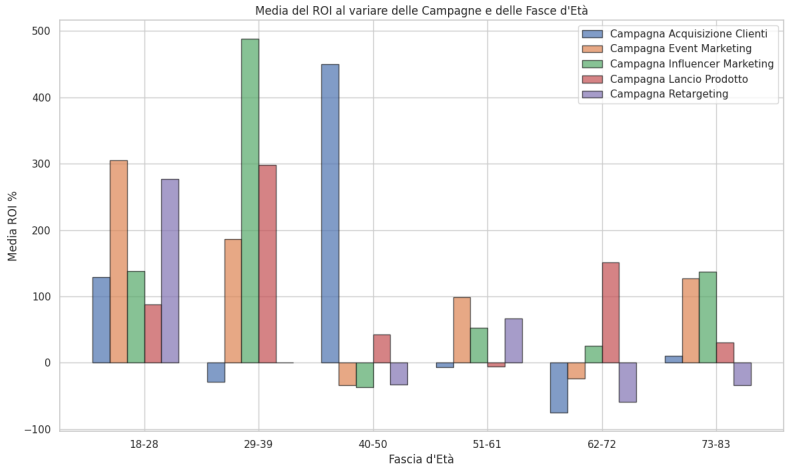

Figura 7.6: Media ROI al variare delle campagne e fasci d'eta'

Esempio Query:

```
df = pd.read_csv('risultato_query.csv')
query = df[(df['profitto'] > df['costo']) & ((df['mese'] == 12) |
    (df['mese'] == 11))]
risultato_finale = query[['ROI', 'campagna', 'tipologia']]
print(risultato_finale)
```

	ROI	campagna	tipologia
17	4.816514	Retargeting	Immagine
24	35.326087	Lancio Prodotto	Testo
41	20.901639	Influencer Marketing	Testo
47	107.675439	Influencer Marketing	Canzone
59	306.578947	Event Marketing	Immagine
93	36.501901	Influencer Marketing	Testo

Figura 7.7: Esempio Query: Profitto maggiore del costo, nei mesi di dicembre o novembre

7.3 Grafici: Sentimento dei Messaggi

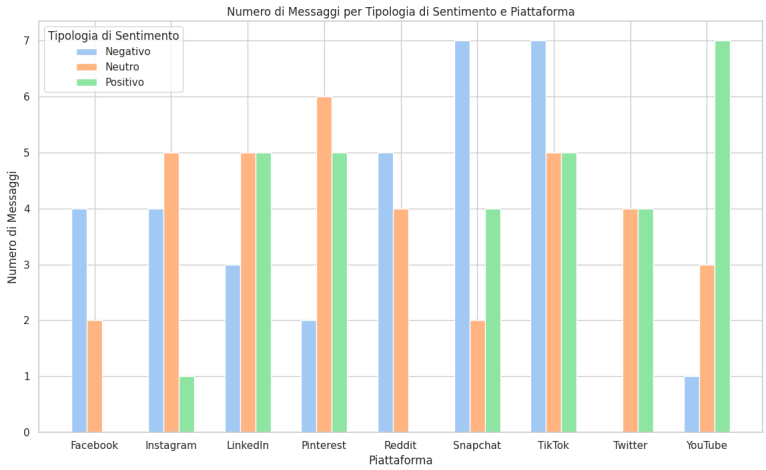

Figura 7.8: Numero messaggi per tipologia di sentimento

Esempio Query:

```
df = pd.read_csv('risultato_query.csv')
query = df[(df['tipologiaSentimento'] == 'Positivo') &
    (df['nome_y'] == 'YouTube')]
risultato_finale = query[['id_utente', 'eta', 'nome_x', 'cognome',
    'demografia']]
print(risultato_finale)
```

	id_utente	età	nome_x	cognome	demografia
5	517	61	Jacob	Becker	West Emilyborough
93	803	20	Carrie	Pearson	Stephanietown
95	60	37	Jennifer	Ibarra	Wilkinsonmouth
96	482	30	Mary	Bowman	Heatherborough
97	918	31	Cody	Graham	South Ericfort
98	448	27	Carolyn	Garcia	Allenview
99	320	71	Andrew	Green	Holmesfort

Figura 7.9: Esempio Query: Informazioni utenti quando la tipologiaSentimento è 'Positivo' su Youtube