POLITECNICO DI TORINO

Collegio di Ingegneria Gestionale

Corso di Laurea Magistrale

in Ingegneria Gestionale

Tesi di Laurea Magistrale

La Business Intelligence e le sue applicazioni in ambito aziendale

Relatore

Prof. Franco Varetto

Candidato

Riccardo La Bella

265168

Aprile 2021

INDICE

Abstract……...1

Introduzione…….……...…….....……..4

Capitolo I

LA BUSINESS INTELLIGENCE……...……..…….……...7

1.1 ***Contesto***……...…….……..…….....7

1.2 ***Cenni storici***……..……...…….……8

1.3 ***I vantaggi della data-driven e della data-oriented***……..…….…….....11

1.3.1 Business intelligence e Business analytics……...……..…….....12

1.3.2 I sistemi informativi…….…….……..…….....13

1.3.3 Database……... ……..…….....18

1.3.4 DBMS……...……... ……..……....……..23

1.4 ***Struttura e architettura di una soluzione BI***……..…….…….....26

1.4.1 Risorse sorgenti.....……...…….……26

1.4.2 Strumenti e processi ETL ……...……..…….....27

1.4.3 Data warehouse ……...……..…….....28

1.4.4 Progettazione concettuale ……...……..…….....31

1.4.5 Progettazione logica ……...……..…….....33

1.4.6 Data mining ……...……..…….....36

1.4.7 Analisi what-if ……...……..…….....47

1.4.8 Data visualization ……...……..…….....49

1.5 ***Software a confronto***……50

Capitolo II

POWER BI…….……..…...……..56

2.1 ***Descrizione*** …….…….……...……56

2.2 ***Struttura e funzioni***……...…….…….....57

2.2.1 Power BI Desktop..…….…….…….…..……...58

2.2.1.1 Connessione ai dati...…….……...……...60

2.2.1.2 Trasformazioni...…….……...……....……63

2.2.1.3 Linguaggio R...…….…….……..……70

2.2.1.4 Linguaggio M...……70

2.2.1.5 Data model...……72

2.2.1.6 Miglioramento delle prestazioni……75

2.2.1.7 Linguaggio DAX...……78

2.2.1.8 Visualizzazione dati...……82

2.2.2 Power BI Service..……97

2.2.2.1 Importazione e caratteristiche principali……98

2.2.2.2 Creazione dashboard...……99

2.2.2.3 Condivisione...……100

2.2.2.4 Pianificazione degli aggiornamenti dei dati……101

Capitolo III

PROGETTO……102

3.1 ***Descrizione*** ……102

3.2 ***Data set***……103

3.3 ***Data editor***……107

3.3.1 Problematiche...……115

3.4 ***Data model***...……117

3.5 ***Relazioni***...……120

3.6 ***Report***……120

3.7 ***Risultati e considerazioni***...……125

3.8 ***Dashboard***...……138

CONCLUSIONI……139

ABSTRACT

Negli ultimi decenni l'innovazione tecnologica ha contribuito in modo esponenziale a facilitare la produzione, la gestione, il potenziamento, la diffusione ed il riutilizzo di tutto quanto sia connesso al patrimonio conoscitivo a disposizione di un'azienda [10]. La costante implementazione delle infrastrutture, finalizzate a creare un ambiente sempre più interconnesso atto ad agevolare la comunicazione tra i dipendenti e la condivisione delle informazioni, lo sviluppo continuo di strumenti via via sempre più ricettivi ed attenti ad un'adeguata considerazione di quanto prodotto, trattato ed utilizzato dalle organizzazioni in termini di conoscenza, l'utilizzo di algoritmi sempre più mirati ed efficaci nella selezione dell'informazione, come anche nella razionalizzazione dei relativi processi di immagazzinamento e recupero dei dati [13], sono tutti elementi che consentono alle organizzazioni di acquisire un costante perfezionamento nella gestione di tale ricchezza [5, 6, 7, 9].

La disponibilità crescente di risorse informatiche sempre più variegate, ad alto potenziale informativo e conoscitivo, di facile e veloce reperibilità (i Big Data) [13], ha accelerato questo processo e indotto le aziende a cercare valore nelle informazioni disponibili all'esterno dei propri confini, in un processo di integrazione dei dati disponibili all'interno con questi nuovi dati esterni [7]. Si tratta di un processo irreversibile ed in costante implementazione, divenuto ormai vitale per la stessa sopravvivenza aziendale, sempre più tesa a fronteggiare la concorrenza ed a mantenere alti livelli di competitività. Ciò ha necessariamente spinto le società a rivedere il ruolo svolto dalla Business Intelligence, arricchendolo di strumenti e procedure nuove e creando ulteriori figure professionali [6].

Obiettivo della presente tesi è quello di fornire una panoramica della business intelligence, della sua origine e della rilevanza che essa riveste in ambito di data-oriented e data-driven, degli strumenti a supporto e della sua utilità in ottica aziendale.

Nel primo capitolo si ha un'iniziale panoramica sulla rilevanza odierna dei dati e del loro utilizzo in ottica aziendale, come debba pertanto essere strutturato un sistema

informativo all'interno di un'organizzazione al fine di massimizzarne i benefici. Si prosegue con la descrizione della Business Intelligence e delle sue applicazioni, in ambito reportistico e decisionale, derivanti dalla comprensione, analisi e condivisione dei dati. Successivamente ne viene spiegata l'architettura attinente ad una sua soluzione nell'ambito di un processo decisionale.

Nel secondo capitolo, dopo il confronto delle diverse proposte sul mercato, si incentra l'attenzione sulle funzioni offerte dal software Power BI realizzato dalla Microsoft, le sue caratteristiche e la sua struttura.

Il terzo ed ultimo capitolo è relativo al progetto previsto dal tirocinio ed all'analisi del caso studio proposto dall'azienda ospitante, prevedente l'estrazione, la rielaborazione e l'analisi dei dati relativi al sito di una azienda cliente; di questo se ne sono valutati diversi dettagli che vanno dagli accessi alle visualizzazioni delle diverse sezioni disponibili all'interno del sito stesso, fino ai tempi di navigazione. Il tutto è stato ottenuto mediante l'utilizzo del servizio "Google Analytics". La trattazione prosegue con l'esplicazione delle modalità di estrazione e di pulitura dei dati e delle tecniche operative di produzione delle informazioni necessarie all'analisi per mezzo di Power BI.

INTRODUZIONE

La struttura della seguente tesi, si è rifatta a quanto affrontato durante il percorso tirocinante svolto. Questo ha inizialmente previsto una fase di introduzione alla Business intelligence con relativi approfondimenti e comprensione delle sue diverse caratteristiche. Successivamente si è passati allo studio del funzionamento del software Power BI per poi incentrarsi, in ultimo, nell'esecuzione del progetto vero e proprio, prevedente l'analisi dei dati inerenti gli accessi al sito di un'azienda rappresentativa di una piccola/media realtà consulenziale, e quelli relativi alla navigazione all'interno dello stesso, operati dai clienti o semplici visitatori. Il tutto, previa estrazione e rielaborazione dei dati mediante l'utilizzo del servizio web di Google Analytics, finalizzate ad ottenere una rappresentazione quanto più possibile univoca dei dati necessari alle funzioni manageriali deputate al processo decisionale per valutare, analizzare e migliorare – se ritenuto necessario - il livello di gradimento, di efficacia e di funzionalità del sito.

Di seguito una breve presentazione della "Certimeter srl" [1, 2, 3, 4], azienda che mi ha ospitato per lo svolgimento del tirocinio e che, con l'occasione, ringrazio per l'opportunità che mi ha offerto, per l'accoglienza riservatami, la disponibilità e la professionalità di tutto il personale col quale ho avuto modo di interagire, con particolare riguardo al tutor che mi ha seguito nel percorso.

Figura 1.

La Certimeter è un'azienda di consulenza informatica specializzata nella realizzazione di soluzioni, prodotti e servizi IT.

E’ la prima società spin-off dell’Università di Torino ed è stata costituita presso il Rettorato dell’Università di Torino nel luglio del 2003, in seguito al deposito ed alla successiva concessione di un brevetto internazionale nel campo delle comunicazioni (USA, Italia, Francia, Germania). Legata a questo brevetto è una delle attività tecnologicamente più caratterizzanti della società Certimeter: l’analisi, lo sviluppo e il test in ambito di sistemi applicativi complessi, networking e sicurezza informatica.

La società, come detto, opera nel campo dell’informatica e delle Reti e, come tale, si pone quale naturale ponte tra il mondo accademico, gli studenti ed i laureati in informatica, le aziende del settore e gli utenti finali dell’ICT.

Certimeter S.r.l. ha affidato una parte delle proprie attività di ricerca e sviluppo al Dipartimento di informatica dell’Università di Torino e ha già ottenuto importanti riconoscimenti, aggiudicandosi ad esempio un finanziamento a fondo perduto di 50.000 euro elargito dalla Provincia di Milano alle migliori imprese innovative (settima impresa su 50 finanziate) e partecipando all’iniziativa “Sinapsi” della regione Piemonte, con un progetto finanziato.

A consolidare la sua posizione nel campo del Web & Network Analytics, la Certimeter è divenuta nel 2005 distributore italiano per il pacchetto software “Sawmill”, leader del settore e particolarmente utilizzato per la varietà delle piattaforme server e di rete supportate. I report di Sawmill sono allineati e integrati con il sistema di certificazione degli accessi proprietario sviluppato dalla Certimeter.

La società è inoltre attiva nello sviluppo software secondo il grame work MVC, seguendo i più accreditati standard in ambiente open source, con applicazioni al settore bancario e assicurativo.

L’azienda è principalmente focalizzata su:

- Software Devolopment & System Integration
- ICT & Cyber Security
- Business Intelligence & Data Analysis

Si avvale della collaborazione di personale altamente qualificato e di comprovata esperienza e capacità: dipendenti, collaboratori e partner formano un network di persone che lavorano sinergicamente per il raggiungimento degli obiettivi e che danno vita ad una realtà aziendale in costante crescita.

La società dispone di due distaccate sedi operative: a Torino in Corso Svizzera n.185 dove è collocata anche la sede legale ed a Milano in Piazza IV Novembre n.4.

Tra i principali clienti serviti annovera:

- Borsa Italiana
- FCA Bank
- Banca Intesa Sanpaolo
- Reale Group
- Deloitte
- CNH
- Arcese
- Juventus
- Reply

La caratura, anche internazionale, delle sopra elencate società clienti, è testimonianza della elevata affidabilità e del livello di competenze e professionalità acquisite dalla Certimeter nei settori in cui opera.

CAPITOLO I

La business intelligence

1.1 Contesto

Al fine di raccogliere, analizzare, condividere ed essere pertanto in grado di gestire in modo efficiente ed al contempo proficuo [11], sia sotto l'aspetto produttivo che economico, la costante crescita delle quantità di informazioni oggi fruibili, sono emerse tecniche, tecnologie, oltre che strumenti, tali da permettere di ricavarne un valido ed oggettivo supporto in ambito decisionale [5], riguardante sia i processi operativi che le scelte in chiave strategica [14].

La business intelligence (BI) si è posta come soluzione irrinunciabile per lo scansionamento, l'archiviazione e l'analisi di contenuti apparentemente grezzi, dai quali estrapolare un preparatorio di conoscenze da convertire, al fine ultimo, in un linguaggio accessibile, di alto livello informativo, in una forma di perspicua comprensione sui diversi aspetti [12], ambienti e finalità di studio da analizzare e interpretare, non consentendone, implicitamente, un'interpretazione basata su fonti pragmatiche o pressoché empiriche, finalizzata da un lato ad una visione di ottimizzazione e miglioramento dei sistemi di interesse e dall'altro ad una riduzione delle loro incertezze [6].

L'analisi delle risorse, sia interne che esterne, in termini più generici dei Big Data [26], risulta quindi ormai essere un utile quanto indispensabile ambito al complessivo ed eterogeneo insieme dei settori di oggigiorno, mostrandosi sempre più frequentemente determinante per il monitoraggio dello stato di funzioni e attività [7], fino a giungere alla moltitudine di scopi e obiettivi che ogni contesto e soggetto si prefigge. L'individuazione di eventuali correlazioni tra i dati significa quindi ottenere un potenziale vantaggio nei corrispettivi rami, un'efficienza della risposta, un solido basamento per le scelte più o meno sensibili [15]. Un loro corretto filtraggio, mediante attinenti tecniche analitiche, comporta di disporre di un'analisi più robusta e di indicazioni più solide [13]. La rilevanza

nel riassumere e nel far coinfluire i contenuti delle risorse in visualizzazioni dalle forme più critiche ed esaustive, rivela l'importanza della salvaguardia e della gestione delle informazioni e del loro corretto utilizzo [14].

La costante proliferazione di tecnologie informatiche sempre più sofisticate ha d'altro canto agevolato la digitalizzazione dei dati, comportando l'implementazione di strumenti ulteriormente più consoni alla loro archiviazione e manipolazione [10]. A questi, si aggiunge il realizzo di modelli e di precise fasi dell'elaborazione conoscitiva che implicano efficienze e risultati differenti a seconda delle diverse congetture adoperate [24]. Una delle principali caratteristiche di base che evidenzia l'effettiva capacità di perseguire gli scopi prefissati dalla BI è l'integrazione delle risorse di diversa origine e struttura [7]; consentirne la possibilità infatti allarga le prospettive di una maggiore ricercatezza e accuratezza della risposta finale [26], passando ovviamente dai precedenti stadi di estrapolazione, ripulitura e gestione delle relazioni [12].

La business intelligence è in grado di rispondere alle esigenze di consapevolezza ma anche in parte di proiezione delle parti e componenti di una specifica realtà, agevolandone le scelte e riducendone il margine di errore. Il continuo sviluppo del settore informativo [11] non potrà che incrementarne il suo peso e utilizzo, rendendolo sempre più necessario ed indispensabile per la gestione di qualsivoglia realtà che operi sul mercato [6, 15, 24, 26].

1.2 Cenni storici

Il primo documento storico [16, 17] che riporta il termine "Business Intelligence" risale al 1865: si tratta dell'opera "*Cyclopaedia of Commercial and Business Anecdotes*" di Richard Millar Devens che coniò tale termine per meglio descrivere il segreto del successo del banchiere Sir Henry Furnese[1]. Quest'ultimo era infatti riuscito ad accumulare ingenti ricchezze raccogliendo, elaborando ed utilizzando informazioni politiche e di mercato, riuscendo così ed operare in anticipo rispetto ai suoi concorrenti.

[1] Mercante e politico inglese (Gunnersbury House, Middlesex (dopo il 1688 - 30 Agosto 1756) e

Nella letteratura moderna il termine “Business Intelligence” nasce ufficialmente nel 1958 ad opera di Hans Peter Luhn[2], con la pubblicazione dell’articolo “*A Business intelligence System*”. Nel novembre di quell’anno, nel corso di una conferenza internazionale dedicata all'informazione scientifica, l'inventore mostrò una serie delle sue macchine elettromeccaniche. Ad una prima osservazione apparvero piuttosto ordinarie, simili agli altri dispositivi informatici già in uso dell'epoca, progettati per raccogliere e ordinare alte pile di schede perforate in slot e contenitori.

Ma a differenza di altri computer, i dispositivi di Luhn non erano progettati per funzionare con numeri e calcoli, ma piuttosto con parole e frasi con l’utilizzo di un algoritmo chiamato da Luhn “KWIC” (Key Word in Context). Prendendo in considerazione un gran numero di testi, in genere articoli da 500 a 5.000 parole, il sistema KWIC era in grado di costruire rapidamente e automaticamente una sorta di indice.

Considerando che all'epoca l'indicizzazione, la classificazione e l'organizzazione delle informazioni scritte era un processo particolarmente lungo e meticoloso, anche per gli specialisti più esperti, e che il volume di informazioni in molti campi cresceva a ritmo incessante, uno strumento veloce e più efficiente che potesse estrarre e riassumere da una gran mole di informazioni, era auspicato ed ormai assolutamente indispensabile. Nel corso di un raduno di bibliotecari e scienziati dell'informazione, tenutosi a Washington, D.C., la dimostrazione del KWIC fu accolta con grande interesse ed entusiasmo, con i giornali di tutti gli Stati Uniti che diedero grande risalto mediatico alla, per l’epoca, sbalorditiva invenzione di Luhn.

Nel processo di sviluppo della B.I. risulta fondamentale l’apporto della IBM, azienda per cui lavorava Luhn, che progetto e diede il via alla produzione dei primi hard disk (1956), dei floppy disk e di altri strumenti di memorizzazione di dati.

Si trattò di strumenti che ben presto si imposero nelle aziende che già si servivano dei computer, consentendo il passaggio dall’archiviazione cartacea a quella digitale, dando

[2] Ricercatore e inventore tedesco (Barmen, 1 Luglio 1896 - 19 agosto 1964)

così anche impulso alla diffusione dei sistemi computerizzati che vennero adottati da un sempre più crescente numero di aziende.

Seguì la creazione del primo Database Management System, ovvero il Decision Support System (DSS). Con i primi anni '70 si assiste ad una crescente competizione tra i progettisti delle diverse aziende produttrici di sistemi operativi, che agevolò lo sviluppo di software sempre più sofisticati e la creazione dei primi *data warehouse*[3] che migliorarono non poco il flusso di dati dai sistemi operazionali ai sistemi di supporto alle decisioni. Il Data Warehousing ebbe il vantaggio di comprimere notevolmente il tempo necessario per l'accesso alle informazioni richieste; ciò ottenuto in funzione del fatto che i dati, sino ad allora memorizzati in diversi database, iniziano ora ad essere a allocati in un'unica postazione.

A questo primo sviluppo si affiancarono altri elementi che oggi stanno alla base della BI come per esempio tool di ETL (Extract, Transform, and Load) e software OLAP (Online Analytical Processing).

Nel ventennio che va dal 1980 ai primi del 2000 la Business Intelligence inizia a diffondersi in misura crescente tra le aziende, anche grazie alla definizione del 1989 data da Howard Dresner [15] , ma il suo pieno utilizzo era ancora ostacolato dalla lentezza nei tempi di elaborazione dei dati e nella difficoltà di utilizzo, ancora riservata a tecnici esperti, i soli in grado di utilizzare software di analisi avanzata. Siamo ancora nell'era della *"Business intelligence 1.0".*

L'ingresso nel ventunesimo secolo è foriero di innovativi sviluppi tecnologici che segnano un vero e proprio punto di svolta e l'ingresso nell'era della "*Business Intelligence 2.0*"; sviluppi che hanno portato a notevoli miglioramenti, sia in termini di complessità che di velocità di elaborazione dei dati. La BI 2.0 si differenzia per l'introduzione di nuove e differenti tecnologie, come ad esempio i processi in real-time, che includevano informazioni provenienti da eventi cosi come apparivano nel Data Warehouse,

[3] William H. Inmon (San Diego, 20 luglio 1945), informatico statunitense che per primo ha parlato esplicitamente di *data warehouse*, lo definisce come una raccolta di dati "*integrata, orientata al soggetto, variabile nel tempo e non volatile*" di supporto ai processi decisionali

permettendo alle aziende di prendere decisioni basandosi sulle informazioni più recenti disponibili. Un'altra tecnologia che venne introdotta includeva accessi self-service per utenti non esperti, questo comportò che gli impiegati potevano ora completare progetti senza dover necessariamente essere supportati o in contatto con l'area IT. Va da sé che ben presto la BI non fu più ritenuta una semplice utility aggiuntiva, o un mero vantaggio, ma stava prepotentemente diventando una necessità per le aziende per continuare ad essere competitive, ponendo quelle che non la utilizzavano a dover rapidamente decidere se adattarsi a loro volta o essere surclassate e messe fuori mercato dai *competitors* più innovativi.

Oggi i tool ed i software della Business Intelligence sono nella maggior parte dei casi *Vertical market software* e quindi tarati per qualsivoglia azienda facente parte di un settore o mercato specifico [15, 17].

1.3 I vantaggi della data-oriented e data-driven

La trasmissione delle informazioni all'interno di una qualsiasi organizzazione è un aspetto che accresce il suo impatto all'incremento della complessità dell'organizzazione stessa. La necessità di coordinamento e compattezza delle informazioni, così come la loro frequenza ed i loro contenuti, rispecchia la struttura del complesso organizzativo, il quale potrà risultare suddiviso mediante differenti tipologie in chiave funzionale o divisionale [18]. A questi si aggiungono ulteriori forme frazionali intermedie, in relazione alle interdipendenze che si vogliono favorire, a seconda del settore e contesto in cui un'azienda, un ente o semplicemente una momentanea collaborazione sia coinvolta [27]. Maggiore risulta la dinamicità dell'ambiente circostante e maggiore sarà la richiesta di flessibilità quanto di coesione necessaria alla comunicazione delle informazioni tra i diversi componenti dell'organizzazione; tutto ciò, infatti, ne condiziona la tempestività delle risposte e l'inerzia alla stoicasticità dei cambiamenti esterni [18].

I dati dai quali trarre le informazioni, il loro modo di essere immagazzinati, la loro manipolazione, l'analisi e la sintesi, divengono pertanto campi sempre più determinanti, proporzionalmente allo sviluppo della digitalizzazione, per ottenere o mantenere un

vantaggio competitivo [15, 18]. I dati si pongono difatti come espedienti strategici nei diversi ambiti di monitoraggio e decisionale, riducendone inversamente un loro utilizzo al solo scopo accessorio [28]. L'informazione terminale, e di conseguenza la scelta intrapresa, risulterà pertanto essere sempre più imperniata sui dati. Proprio il concetto di data-orieted o data-driven indica di conseguenza il voler improntare la direzione di un apparato sulla fruizione dei dati, al fine di vincolarsi all'oggettività ed empiricità che questi rappresentano [20]. La pianificazione relazionata ai dati disporrà quindi di un supporto costante alle innumerevoli considerazioni necessarie per il perseguimento degli obiettivi e dei risultati attesi nei differenti strati dell'organizzazione, alla base di ogni progetto ultimo [21].

1.3.1 Business intelligence e business analytics

Nel quadro descrittivo delle strategie della data-driven, è possibile differenziare i diversi ed effettivi utilizzi dell'informazione elaborata dai dati [21]. Mediante specifici sistemi e tecniche è consentito proiettarsi, dai medesimi o meno aggregati di dati, sia ad una visione passata ed implicitamente descrittiva di un insieme di istanti (o unico) della realtà a noi di interesse, sia ad un contesto temporale futuro, con intento predittivo e prescrittivo [23]. La determinazione delle ramificazioni dell'uso dell'informazione [5] consente a sua volta di distinguere, in modo analogo, le differenze tra la business intelligence e la business analytics. Seppur in alcune parti sovrapponibili [35], la BI e la BA differiscono per il periodo temporale al quale si rivolgono e, corrispondentemente, ai presupposti e alle domande che vengono poste, delle quali se ne ricercano le constatazioni e le risposte [29, 30].

La business analytics cerca di anticipare eventi e situazioni, di spiegare specifici eventi, determinare nuovi modelli e correlazioni [23]; prima di effettuare aggregazioni analizza nel dettaglio i dati [36], applica formule statistiche e solo dopo trova correlazioni e modelli in grado di realizzare delle previsioni. La business analytcs persegue tali scopi mediante diversi strumenti e tecniche quali l'econometria, l'analisi di serie e l'analisi statistica inclusi l'albero decisionale e la regressione logistica, l'analisi delle deviazioni ed il rilevamento delle anomalie, l'ottimizzazione, la simulazione, l'analisi testuale e l'analisi

quantitativa [31]. Utilizzata principalmente nel marketing e nella previsione dei comportamenti del cliente [33], si è in seguito espansa nella determinazione di scenari riguardanti le aziende stesse in relazione a mutamenti sia interni che esterni [25, 34].

La business intelligence si focalizza invece sulla valutazione delle serie storiche e dell'attimo presente [23], evidenziandone i trend e gli aspetti, così da realizzarne benchmark attinenti. Si rifà principalmente al reporting come i KPI, grafi e misure, al monitoraggio e quindi all'analisi descrittiva e diagnostica, dashboard e analisi delle relazioni multidimensionale (OLAP) [32, 33]. In comune alla business analytics dispone necessariamente della collezione e gestione dei dati, del data mining oltre che le presentazioni e le visualizzazioni stesse risultandone pertanto parte integrante [33].

Entrambe, insieme, riescono di conseguenza a coprire l'intera valutazione temporale seppur con relative incertezze e criticità. Contribuiscono in egual modo all'efficienza delle scelte decisionali in contesti strategici, tattici e operazionali in quanto, seppure maneggiati in modo differente, si basano su dati non soggettivi [22, 25].

1.3.2 I sistemi informativi

Un'organizzazione, ovvero un gruppo di soggetti ed entità fra loro in interazione, a seconda dei processi interni corrispondenti alle funzioni, agli scopi, al contesto con il quale interagisce, si dispone in differenti livelli ed ambiti che necessitano, per la loro comprensione e gestione, di determinate e specifiche informazioni a riguardo, oltre ovviamente che di relative conoscenze [18]. Maggiore è infatti la complessità dell'organizzazione, in dimensione e segmentazione, e maggiore sarà il rigore verso il dettaglio e la compattezza che le informazioni veicolate dovranno disporre, attinenti ad ogni preciso settore. Di quest'ultimo, l'accuratezza dei dati consente quindi di acquisirne la facoltà di comprensione dello stato e difatti ne comporta la corretta o meno deduzione delle conclusioni, il coordinamento, lo sviluppo, il giusto supporto alle scelte da intraprendere. Alla base quindi della comunicazione delle informazioni all'interno di un apparato organizzativo, vi sono in conclusione un insieme di relazioni tra i medesimi soggetti dell'organizzazione stessa, i suoi processi, le sue funzioni e la sua struttura, che

costituiscono nel complesso i cosiddetti sistemi informativi [37, 39]. Dopo pertanto la deduzione degli orizzonti temporali ai quali può l'informazione riferirsi (mediante la distinzione della Business Analytics e della Business Intelligence) che determina già quali dati e tecniche di analisi siano necessari, le complessità ulteriori per la composizione di un sistema informativo adeguato e ben strutturato sono sia la stretta dipendenza al personale a cui è rivolto e che dovrà materialmente estrapolare e decidere, sia la complessiva intelaiatura dell'apparato circostante al medesimo destinatario, la quale sarà più o meno direttamente o implicitamente coinvolta [38]. In ausilio alla risoluzione di quest'ultime complicanze, si sfruttano diversi strumenti e alternative che consentono di identificare e classificare i meccanismi dediti alla realizzazione e veicolazione dell'informazione. Al perseguimento di questo intento, si definisce per primo l'ambiente in cui agisce il sistema informativo e gli attori interni ed esterni oltre che le influenze. Si definiscono di conseguenza le fasi e quindi gli elementi costituenti, come vengono utilizzati e come sono individuati.

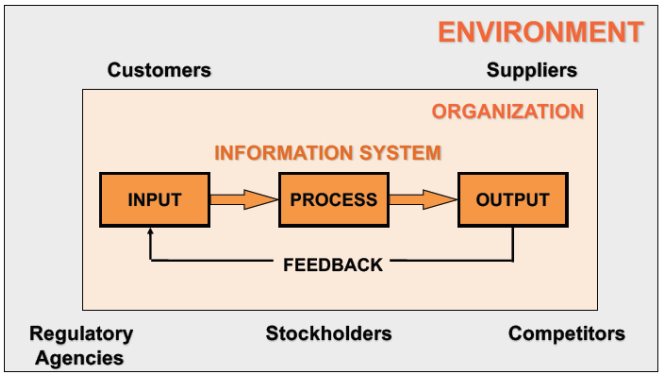


Figura 2.

L'organizzazione e l'ambiente nel quale il sistema informativo opera [18, 40, 43]:

- **Risorse umane** = manager, addetti alla produzione o ai servizi.
- **Risorse tangibili** = tecnologie, locazione, impianti, brevetti.
- **Risorse intagibili** = capacità, competenze, cultura, esperienza, reputazione, fidelizzazione.

- **Processo di business** = organizzazione delle attività
- **Struttura** = organigramma, geografia, per gruppi di specialisti e/o di prodotti.
- **Funzione aziendale** = compiti e obiettivi specifici che l'organizzazione si pone di raggiungere in relazione al consumatore a cui si rivolge, alle proprie caratteristiche, ai fornitori.
- **Fattori esterni** = cliente, fornitore, concorrenti, azionisti, agenzie regolamentarie

All'interno di questo articolato contesto, il sistema dovrà porsi come collante tra le diverse dimensioni. Nel particolare dovrà risultare maggiormente compatto in corrispondenza delle cosiddette competenze core, ovvero le relazioni tra i diversi elementi che permettono di conferire un maggior vantaggio competitivo all'organizzazione: sostanzialmente le competenze cardini determinanti nell'efficienza dei prodotti e/o servizi offerti. Al fine di determinare le competenze core, si utilizzano diversi strumenti quale l'analisi VRIO[4] che predispone le basi identificative sulla presenza di effettive correlazioni tra le risorse e la capacità, tali da generare un rilevante valore aggiunto. Analogamente, è necessario che ci siano forti legami in chiave di processi, struttura organizzativa, motivazione, consistenza e allineamento. Nella maggior parte dei casi, si ha un ulteriore supporto mediante altri mezzi come la "catena del valore" e la "catena dei fornitori", decisivi per la comprensione dei processi e pertanto delle attività che li costituiscono e che più coinvolgono i sistemi informativi [18].

Negli ambiti di carattere maggiormente strategico e meno operativo, invece, corrispondenti alla funzione aziendale, si dispone ad esempio del "diamante di Porter" per lo studio competitivo in una determinata area geografica. Si aggiungono le analisi macroeconomiche di un settore necessarie per delinearne i confini, la segmentazione del mercato ed i fattori chiave. Il tutto dovrà essere conforme alla struttura, principalmente indentificabile nelle macroaree della sezione operativa, della strategica

[4] Analisi sviluppata da Jay Barney in " Risorse aziendali e vantaggio competitivo sostenibile (1991)". L'acronimo VRIO sta per: "V = valore", "R = rarità", "I = inimitabilità" e " O = organizzazione"

e della gestionale (denominata anche middle line) oltre che della potenziale presenza dello staff di supporto e della struttura tecnologica [18].

Fasi e parti costituenti di un sistema operativo:

- **Input** = acquisizione dei dati grezzi dall'interno dell'organizzazione o dall'ambiente esterno. I dati sono di per sé privi di significato in quanto risultati empirici di determinate circostanze della realtà.
- **Processing** = manipolazione ed elaborazione dei dati grezzi al fine di individuarne ed estrapolarne correlazioni da convertire in un linguaggio di alto livello, in informazione, accessibile e quindi significativa.
- **Output** = distribuzione della conoscenza ricavata dall'apprendimento al fine che venga interpretata dal personale e utilizzata per le diverse attività.
- **Feedback** = valutazioni provenienti dall'applicazione della conoscenza e perfezionate dall'esperienza che andranno a proporre correzioni alle fasi antecedenti.

L'essenza dei sistemi operativi, difatti, risulta essere costituita dalla raccolta dei dati da un livello inferiore o meno dell'organizzazione (oppure dall'esterno, in relazione alla risorse coinvolte e settore di appartenenza), la loro gestione, la loro sintesi ed infine la trasmissione ai livelli più alti. Il tutto mantenendo la coerenza e la stretta vicinanza ai processi e competenze coinvolti. Solo dopo aver delineato e descritto il quadro completo di rappresentazione e di azione dei sistemi operativi, è pertanto possibile poterli classificare a seconda di due diverse tecniche: per mezzo della suddivisione dell'organizzazione in livelli e funzioni (maggiormente dettagliata) oppure tramite l'aggregazione in famiglie [40, 41].

Per la prima tecnica, si utilizza la cosiddetta "piramide di Anthony" che definisce due dimensioni nelle quali è immersa l'organizzazione: ovvero il livello organizzativo e la funzione aziendale.

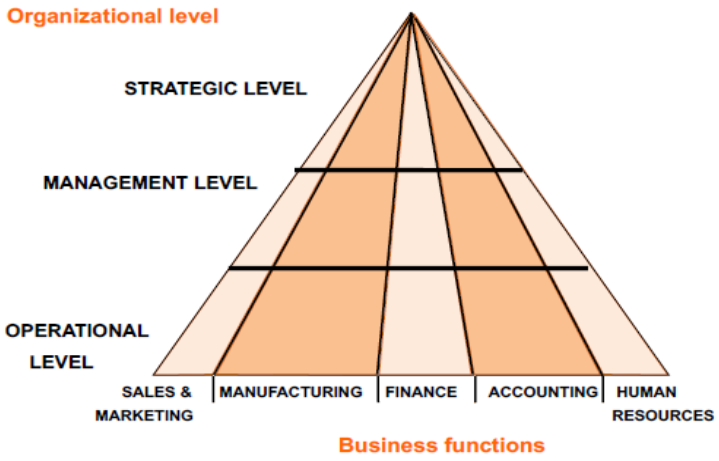


Figura 3.

La prima dimensione identifica tre livelli:

- **Operativo** = corrisponde all'insieme delle attività maggiormente coinvolte all'interno dei processi dell'organizzazione. Sono qui presenti una elevata quantità di dati, oltre che a un non indifferente tasso di operazioni. I relativi SI in questo livello prendono il nome di Transaction Processing System (TPS) per la gestione delle transazioni.
- **Gestionale** = riguarda le attività attinenti alla gestione dei processi nel livello inferiore, ovvero l'operativo. Il tasso di uso è parallelamente inferiore così come la quantità di dati adoperati. Emergono i primi KPI.
- **Strategico** = coinvolge le attività di strategia, meno frequenti rispetto ai livelli inferiori e bisognosi di un quantitivo di dati ancora ampio ma più compatto, sintetizzato.

La seconda dimensione rappresenta invece le funzioni aziendali. I riquadri ottenuti dall'intersezione tra i livelli e le funzioni, corrispondono a determinate caratteristiche necessarie per un sistema informativo, il tutto al fine di ottenere uno strumento che possa maggiormente permettere il raggiungimento di un sufficiente grado di efficienza e prestazioni, a seconda dello specifico ambito [40, 41, 42, 45, 49, 51].

Nel dettaglio, per il livello strategico si adoperano i SI ESS (Executive support system) che producono nel particolare proiezioni utili ai senior manager. Nel livello gestionale se ne differenziano di diversi come i MIS (Management information system), i DSS (Decision support system), KWS (Knowledge work system) e OAS (Office automation systems). Questi supportano in ambito organizzativo, riuscendo a produrre informazioni aggregate ed a realizzare simulazioni. A livello operativo si hanno invece i TPS (Transaction processing system) che coinvolgono operazioni semplici ma di numero estremamente elevato [40, 41, 50] .

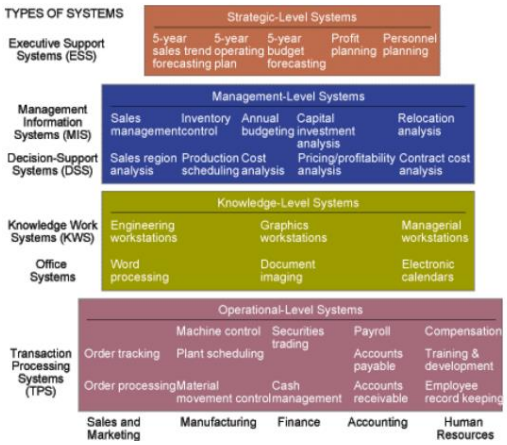


Figura 4.

La tecnica fondata sulla categorizzazione dei sistemi informativi, invece, si basa sul realizzo di famiglie che possono essere utilizzate in relazione ad esigenze meno dettagliate rispetto alla precedente tecnica; nella pratica, a seconda della funzione aziendale di appartenenza (settore manufatturiero, bancario etc) si adoperano sistemi informativi già predefiniti per quel preciso contesto, che non andranno a colmare peculiari aspetti dell'organizzazione, ma risulteranno sufficienti per procedure generali [40, 41].

1.3.3. I database

Indispensabile, per quanto descritto nei paragrafi precedenti, risultano conseguentemente essere le abilità riguardo l'archiviazione dei dati ed il loro stoccaggio. Di questi, infatti, se ne incrementano costantemente i volumi, comportandone un

continuo e crescente bisogno di capacità sia fisiche quanto conoscitive [52, 53]. Ad esempio, sempre più frequentemente, in molte realtà non è del tutto sufficiente disporre di una sola macchina come spazio a disposizione per la memorizzazione e, per le dimensioni e per la mole di dati oggi reperibili e disponbili, come detto, si è sempre più obbligati a dover ricorrere ad un cluster di macchine. Al volume si aggiungono però altre importanti caratteristiche e aspetti non meno rilevanti, imprescindibili per il raggiungimento minino di efficienza: la velocità di memorizzazione, la varietà di dati da dover essere capaci di gestire, la loro veridicità [60].

In questo contesto informativo si avranno pertanto le cosiddette basi di dati, ovvero gli strumenti dediti alla collezione delle informazioni, sostanzialmente strumenti congeniali alla configurazione dei dati stessi, ed i DBMS (Data Base Managament System) cioè gli specifici sistemi software che ne consentono la gestione [56, 59].

Le banche dati sono quindi un elemento cardine per i processi di analisi e di monitoraggio. Da questi si estrapolano informazioni compatte e sintetiche, rilevanti per il mantenimento e lo sviluppo dei singoli processi così come che dell'intero apparato. Occorre però conoscere quanto più nel dettaglio le diverse congetture a disposizione per il loro realizzo, al fine di sfruttarne al meglio le efficienze e incrementarne le prestazioni. Al di là degli aspetti prettamente più tecnici, riguardanti prevalentemente l'assetto fisico, vi sono due non trascurabili approcci da tenere in considerazione per essere in grado di delineare correttamente prima e gestire dopo, l'informazione all'interno. Nello specifico riguardano l'ambito concettuale da un lato e l'ambito logico dall'altro, le cui diverse possibilità e alternative di impiego ed interazione conseguono a differenti soluzioni [59]. Il modello concettuale corrisponderebbe alla rappresentazione della realtà, dei dati quindi a disposizione, in modo indipendente dalle diverse tipologie dei modelli logici.

Sulla tipologia relativa alla struttura, e quindi all'ambito logico invece, ci si riferisce nel dettaglio alle modalità rappresentative dell'informazione, ossia al formato dei dati elementari, il cosiddetto "type" di appartenenza (come potrebbe ad esempio risultare la classificazione in numero intero o decimale). In questi modelli rientrano comunque

anche i diversi criteri logici da seguire nel realizzo della configurazione, per indirizzare l'intero complesso verso strutture più o meno complesse, come possono essere gli array [59].

Oltre a questo aspetto, risulta di pari importanza nelle considerazioni per la costituzione delle basi di dati, soprattutto le connessioni tra le informazioni che verranno contenute all'interno delle basi di dati, delineate in fase di progettazione nel concettuale e messe in pratica nei modelli di tipo logico. Questi ultimi risultano essere numerosi e abbastanza eterogenei l'uno con l'altro ma il più utilizzato, e pertanto diffuso, è senz'altro il modello relazionale [59, 60]. Il motivo, molto probabilmente, è anche dovuto alla sua praticità che ne agevola una sua più rapida comprensione. Il modello relazionale infatti, altro non si rifà che all'utilizzo di tabelle, il cui scopo è quello di raggruppare per costrutto un insieme omogeneo di record [59, 52]. Vi sono, come accennato, altri modelli logici sia precedenti che successivi temporalmente al relazionale, altri ancora si pongono come sue sfaccettature. Tra i più conosciuti ci sono i modelli NoSQL che ne comprendono diverse tipologie a secondo che l'archivazione sia su base colonnare, di grafi, documentale o di coppie valori-chiave [60]. Si differenziano dal relazione per la capacità di immagazzinare dati non strutturati o semistrutturati, ma anche di salvare dataset di grandi dimensioni con scalabilità e prestazioni più economiche. Inizialmente furono utilizzati i modelli di tipo gerarchico e reticolare che sono meno astratti e più uniformi alla struttura fisica stessa del database. Altri ancora sono il modello ad oggetti, ai quali se ne aggiunsero di ulteriori come i data warehouse, i database OLTP, i distribuiti, i cloud e open source [56]. Ognuno di questi modelli ha preso ovviamente piede per ragioni differenti, a seconda della soluzione in quel momento più congeniale da intraprendere, riguardante i diversi bisogni degli specifici ambiti coinvolti [79]. Il modello relazione si è imposto a partire dagli anni '80 riuscendo a mantenere la sua egemonia per i successivi decenni [53]. E' una soluzione in grado, e lo è tutt'ora, di fornire un meccanismo di archivazione di tipo persistente dei dati oltre che di un controllo della concorrenza e delle transazioni, riuscendo a disporre di interfacce standard, sia per l'integrazione che per la creazione di report. I database relazionali convenzionali, utilizzano tabelle

bidimensionali per incorporare i dati [53], con una solida affidabilità delle operazioni. Le tabelle sono definite mediante la delineazione dei cosiddetti “schemi”, ovvero le semplici intestazioni e denominazioni delle colonne; principalmente non variano nel tempo e si prestano a tracciare la struttura ad aggregazione dei dati. A rappresentare la seconda dimensione sono invece le “istanze”, concetto riconducibile ai contenuti effettivi all’interno delle righe delle tabelle stesse [59]. Le istanze possono, a differenza degli schemi, variare nel tempo e sono l’oggetto dell’estrapolazione per la creazione dell’informazione. Il linguaggio associato al modello relazionale è il linguaggio SQL [52], che si predispone di due diverse macrocategorie relativi ai comandi, in relazione al loro fine ultimo. Da questi infatti si distinguono i linguaggi di definizione dei dati (Data definition language o DDL) adoperati per delineare gli schemi logici, interni ed esterni, ciò che riguarda pertanto la creazione del dataset e parallelamente i linguaggi di manipolazione dei dati (Data manipulation language o DML) intrapresi per l’aggiornamento, l’estrapolazione e quindi l’interrogazione delle istanze [68]. Il modello relazionale è l’approccio più in uso nel contesto dei database ma ciò non perchè sia totalmente esente da criticità, alle volte anche di non trascurabile impatto. Il modello mostra infatti debolezze in diverse circostanze, relative nella maggior parte alle operazioni di lettura e scrittura dei dati in tempo reale; ciò avviene principalmente nelll’esecuzione di query SQL composte, nello specifico se correlate ad un numero elevato di tabelle che, inevitabilmente, ne riducono le prestazioni in termini di tempo, di risorse utilizzate e facilità della comprensione per i meno esperti [52, 53]. Altre difficoltà si riscontrano per via della stretta dipendenza ad uno schema archivazionale rigido che complica il realizzo di nuove relazioni tra le entità [53].

A fine anni ‘90 si realizzato un database relazionale open-source, privo di un’interfaccia SQL e che adoperava un numero ridotto di risorse [60]. Circa dieci anni più tardi i database non relazionali cominciarono a rafforzare le proprie fondamenta concettuali, con riscontri nettamente positivi da parte degli utenti; da quel momento furono utilizzati con sempre maggior continuità e inziarono ad essere accostati alla denominazione dei modelli “NoSQL”, proprio per l’assenza dello specifico linguaggio “Structured Query

Language" [60]. Le differenze con il modello relazionale sono molteplici e ricoprono diversi aspetti. Come accennato, tali modelli si rifanno anzitutto a linguaggi personalizzati, distanti dallo SQL, incentrati sui concetti di documenti, grafici e altre strutture specializzate. Non vi è dipendenza da alcuno schema, nessuna tabella pertanto, ma si fondano sulle diverse forme di archiviazione su base colonnale o per coppia chiave-valore; gli schemi pertanto possono variare a seconda del caso. Non sono presenti operazioni di "join" [53], di giunzione, il che riduce considerevolmente la complessità delle query consentendo l'incremento, parallelamente, delle prestazioni. Con i modelli non relazionali è possibile infatti archiviare ed elaborare grandi volumi di dati [53]. Altra caratteristica nevralgica è la scalabilità orizzontale ovvero la possibilità di ridurre il carico dei database introducendo nuovi server. I modelli NoSQL sono i più conformi per il trattamento di dati non strutturati simili a XML e JSON agevolandone ulterioremente il loro uso [60]. Tra i modelli non relazionali si hanno [54, 55]:

- **Chiave-valore**
- **Orientati per colonna**
- **Database su grafi**
- **Basati su documenti**

Altre tipologie di database sono i "distribuiti" [60] che possono o meno essere interamente non relazionali. Sono caratterizzati da una ridondanza di dati collocati in locazioni differenti. La loro introduzione è stata dovuta principalmente, oltre che per il miglioramento delle performance, all'esigenza di trovare delle soluzioni in caso di guasti e di poter disporre di diverse copie dei medesimi dati; questo consente di non perdere in nessun caso l'informazione contenuta e conservata, anche nella circostanza in cui una o più macchine dovessero interrompere il proprio servizio. Le possibilità a tale proposito sono diverse, ovvero la replicazione di tipo Master-Slave, di tipo sincrone oppure asincrone. Queste offrono rimedi differenti seppur condividendo per la maggior parte gli stessi assetti ed elementi concettuali. I database distribuiti sono oggetto del cosiddetto teorema CAP il quale esplica come vi sia l'impossibilità che vengano soddisfatti contemporaneamente la caratteristiche di consistenza, disponibilità e tolleranza al

partizionamento. Da questa denotazione si sono proposte alternative più deboli quali la BASE (Basically Available, Soft State, Eventually Consistent) differente dalla ACID (Atomicity, Consistency, Isolation, Durability) più complicata invece da perseguire nel contesto dei database distribuiti [60].

1.3.4 I DBMS

Al fine di gestire e organizzare, dirigerne l'accesso ed il controllo delle collezioni di dati, ovvero delle basi di dati, si adoperano determinati software; questi prendono il nome di DBMS (Data Base Managemente System) e si pongono come servizi necessari quanto indispensabili per il caricamento e l'aggiornamento dei dati influenzandone la struttura stessa dell'archivazione [72, 74, 78]. I DBMS consentono di ridurre l'inconsistenza dei dati così come la loro ridondanza [59] mediante un meccanismo di controllo dell'accessibilità di tipo concorrenziale. Le banche dati infatti risultano essere delle risorse integrate, condivise fra più reparti, applicazioni e ambiti della medesima organizzazione. I database management system non si pongono però come garanti della sola gerenza di condivisione della mole di dati, ma ne assicurano anche la privatezza, l'affidabilità, la persistenza, l'efficienza e l'efficacia [68]. La permanenza dei dati perdura difatti anche al termine dell'esecuzione di altri software che vi lavorano per differenti scopi; le autorizzazioni per l'accesso coinvolgono robusti sistemi in back-end e nell'eventualità di disservizi dovuti a complicanze, in ambito sia hardware così come attinente allo svolgimento di programmi, si attuano il ripristino e quindi i recovery delle ultime memorizzazioni, pertanto dei backup [70]. A queste caratteristiche, i DBMS forniscono ulteriori sostegni, quali l'ottimizzazione delle prestazioni mediante una riconsiderazione delle effettive dimensioni degli apparati informativi in temini di risorse e durata; coloro che fruiscono delle competenze a disposizione sono inoltre in grado di massimizzarne l'utilità ricavabile, anche grazie alla facilità d'uso della maggior parte dei prodotti offerti. Si ha oltretutto un incremento delle funzioni disponibili in comparazione a quanto proposto dai files system in termini di salvataggio e interrogazione dei dati, di meccanismi di lettura (vincolati alle sole sequenziali) e di condivisione. La struttura dei DBMS è delineata mediante gli schemi logici, dai quali ne sono riconducibili ulteriori

manifestazioni in dipendenza delle destinazioni, se interne o esterne, a seconda a loro volta che si descrivano rispettivamente gli schemi logici sulla base delle strutture fisiche di memorizzazione o se invece sulle "viste" lato utenti [68]. Da questa separazione in livelli di astrazione è possibile riscontrare la non stretta dipendenza dei dati stessi tra le risorse tangibili di archivazione, i database pertanto, ed i sistemi applicativi adoperati. I database management system si collocano in una posizione intermedia fra questi, tra il livello fisico e la rappresentazione logica [77]. Gli utenti che usufruiscono dei servizi visualizzano, infatti, le relazioni tra i dati in considerazione della sola struttura delineata nel livello logico e proposta nel DBMS, indipendentemente e separata dalla composizione del livello più basso ed effettivo pertanto, costituito dai server, dischi e ulteriori macchine o risorse; l'accesso avviene difatti nel medesimo modo a prescindere delle diverse e mutevoli scelte di memorizzazione fisica dei dati, senza influire pertanto sui programmi per la loro gestione [77]. I DBMS sono sistemi che coinvolgono d'altro canto molteplici figure, ognuna della quali con ruoli e funzioni diverse. Una delle maggiori per responsabilità, è il database administrator su cui è incentrata la gerenza ed il controllo delle basi di dati; a lui sono associate le funzioni per la disciplina delle autorizzazioni all'accesso, oltre che risultarne garante dell'affidabilità e del raggiungimento minimo delle prestazioni. Altre figure sono rappresentate dai programmatori e dai progettisti, incaricati dell'architettura delle banche di dati e della programmazione dei sistemi applicativi che ne dispongono l'utilizzo. Per ultimi si hanno gli utenti che usufruiscono dei servizi per le proprie attività; quest'ultime prendono il nome di transazioni se predefinite e di conseguenza già comprese in specifici programmi, mentre sono ritenute casuali le semplici interrogazioni [59].

I database management system sono software complessi quanto costosi che richiedono un personale preparato e specifiche risorse in ambito informatico per poter massimizzarne le prestazioni. I vantaggi sono però elevati in quanto si ha modo di ottenere non trascurabili economie di scala per mezzo della centralizzazione e standardizzazione dei dati [59]. Differenti applicazioni possono essere a questi riconducibili inoltre, ottenendo numerevoli alternative e possibilità a seconda

dell'esigenza d'uso degli utenti. In egual modo anche i DBMS stessi possono distinguersi a seconda del fine ultimo dei relativi dati a disposizione all'interno dei database. Nell'ambito della business intelligence occorre nel particolare distinguere le divergenze tra due tipologie di DBMS, rispettivamente con tecniche OLAP ed OLTP.

- **I sistemi OLTP (On-Line Transaction Processing)** sono accostati comunemente ai database di tipo tradizionale riproponendo le sue proprietà ACID. L'utilizzo di questi sistemi è principalmente finalizzato alla sole transazioni ed elaborazione delle query: pertanto ad operazioni di tipo standard, rappresentate dalla lettura dei dati o al massimo da loro modeste modifiche. Sono adoperati da operatori ed il loro utilizzo è frequente. I dati risultano essere di tipo corrente, dettagliati e aggiornati, senza riferimenti a dati storici o dati di diversi apparati in quanto l'interesse è incentrato alla attività quotidiane, operazionali: non occorre quindi coprire larghi spazi temporali. Con riguardo alla progettazione, invece, i sistemi OLTP adoperano il modello ER entità-relazione mentre lo sviluppo del database è orientato alla applicazione [76]. Le modalità di accesso invece richiedono meccanismi di controllo, ripristino e concorrenza in quanto sarà di norma utilizzato contemporaneamente da un elevato numero di utenti. L'intento principale di questi sistemi è proporre soluzioni predefinite con poco margine di errore, in modo tale da velocizzarne le rispettive operazioni [67, 75].
- **I sistemi OLAP (On-Line Analytical Processing)** contribuiscono invece all'analisi di dati al fine di delineare le differenti constatazioni e opzioni, consentendo di fatto una scelta più strutturata. I processi decisionali fanno uso di un aggregato di dati sintetizzazi e riassunti, che coprono un arco di tempo sufficientemente ampio da poterne evidenziare le variazioni. Non a caso, i sistemi OLAP si differenziano dai sistemi OLTP proprio per l'uso finale, dato che non vi è l'esigenza di gestire ordinarie operazioni bensì di assumere direttive quando risulta effettivamente necessario. In questo contesto, i sistemi OLAP sono pertanto usati dai dirigenti, analisti e manager e indirizzati a scopi attinenti al mercato. I dati, in quanto l'arco temporale è come detto più ampio, saranno particolarmente numerosi, ma

archiviati e gestiti su diversi livelli al fine di garantirne una più facile estrapolazione a secondo degli aspetti decisionali da adottare [76]. L'uso è pertanto ad hoc mentre l'accesso è principalmente basato alla sola lettura sebbene vi siano query estremamente complesse [75]. I modelli meno utilizzati per la progettazione dei database, invece, riguardano i modelli a stella o a fiocco di neve che consentono di organizzare la struttura dati in modo più astratto [67].

1.4 Struttura e architettura di una soluzione BI

L'elevato quantitativo di dati oggi disponibili nei database gestiti dai DBMS, contribuisce alla potenziale opportunità di ricavare informazioni utili sul piano decisionale e gestionale riguardo ad ogni parte costituente di un'organizzazione e non solo. L'analisi dei dati permette di descrivere situazioni attuali, ma anche di proiettare eventuali futuri scenari. La complessità si riscontra però nel disporre delle corrette e adeguate conoscenze, oltre che strumenti, capaci di generare valore dai sistemi informativi in correlazione alla struttura dell'apparato organizzativo stesso. La business intelligence si pone come soluzione, collocandosi come collante tra i differenti elementi e livelli informativi, riuscendo ad estrapolare dai dati apparentemente grezzi informazioni che possano conferire un vantaggio competitivo [15]. A tal fine si avvale di un insieme di parti e congetture.

1.4.1 Risorse sorgenti

Corrispondono alle fonti dalle quali rifornirsi di dati grezzi. Possono provenire dall'interno o dall'esterno dell'orgarnizzazione [81] e comprendono diversi formati e sistemi. Solitamente vengono forniti dalla Custumer Relationship Management (CRM) e Content Management (CMS), dalla Enterprise Resourse Planning (ERP), dalla Supply Chain Managemente (SCM), dall'ecommerce ma anche da fornitori esterni, WEB, applicazioni, servizi cloud etc [85, 86].

Le forme più diffuse sono per la maggior parte tabulazioni fornite da database e DBMS; i flat file come Excel, CVS, file di testo, JSON e XML sono anch'essi origini frequenti [81,

85]. La provenienza ed i formati variano a seconda dell'ambito di interesse dell'organizzazione.

La varietà delle risorse a disposizione da trarre come riferimento alle successive analisi, è la constatazione di quanto sia stata determinante l'esigenza di risultare il maggior possibile aperti ed eclettici alle nuove alternative, sviluppatesi con la costante crescita di informazioni sempre più ricche ed accessibili [119].

1.4.2 Strumenti e processi ETL

Una volta identificate le risorse con le loro varianti e differenze, occorre inoltrarsi nelle rispettive procedure di estrazione, trasformazione e caricamento dei dati che nell'insieme vengono denominati come processi ETL (Extraction, Trasformation and Loading) [82]. Questi, appartenenti al back-end, consentono mediante appositi strumenti e procedure, di uniformare i dati delle diverse sorgenti, affinchè vi sia omogeneità, correttezza, coerenza e robustezza di ciò che sarà il contenuto di un archivio in via di popolamento o aggiornamento, quale può risultare il data warehouse.

E' una parte cardine per la realizzazione dell'informazione finale; secondo Ralph Kimball e Joe Caserta i processi ETL possono giungere ad impegnare circa il 70% di tutte le risorse necessarie per il realizzo del data warehouse [82]. La loro rilevanza cresce infatti all'aumento della complessità delle fonti, delle interdipendenze dei dati oltre che della loro qualità e dello spazio di archivazione a disposizione [67, 61, 84, 85].

- **Estrazione:** nella prima fase vengono estrapolati i dati, principalmente dai sistemi operativi per di più fra loro eterogenei. Qui viene ridotta la quantità di dati omettendo qualsiasi set non prettamente rilevante. Nella maggior parte dei casi avviene in background o eseguita nei momenti di scarsa attività, quali ad esempio le ore notturne. Questo al fine di non influire negativamente sulle prestazioni dei sistemi produttivi. L'aggiornamento periodico prende il nome di estrazione incrementale.
- **Pulitura**: successiva all'estrazione, è un'operazione volta a migliorare la qualità dei dati in correttezza e consistenza. Si va pertanto a rimuovere dati duplicati,

errati o impossibili, mancanti; si correggono eventuali usi non previsti di un campo o l'inconsistenza tra valori; oltre a questi aspetti, si procede inoltre all'eliminazione di errori di battitura e incongruenza nei formati.

- **Traformazione:** in questa terza fase si ha invece la conversione dei dati al fine che possano essere interpretati in termini di business; ciò avviene in relazione della qualità e correlazione dei e tra i dati stessi. Successivamente vengono uniformati nello schema del database di destinazione, integrati e consolidati.
- **Caricamento:** quarta e ultima fase, si inseriscono i dati all'interno del data warehouse. Questo passaggio può avvenire al fine di popolare per la prima volta il database o aggiornarlo.

Un ulteriore elemento, a supporto dei processi ETL, è la **"Staging Area"** o **"Operational Data Store"** (ODS) che può o meno essere adoperata a seconda delle differenti architetture intraprese [83, 61]. È un'area di transito che permette di separare le prime tre fasi dalla quarta e pertanto l'estrazione, la pulitura e la trasformazione dal caricamento [85]. Questa scelta permette di intraprendere operazioni più elaborate e complesse riguardanti la pulizia e la conversione dei dati, vicine alla rappresentazione OLTP; genera d'altro canto ulteriore ridondanza provocando il bisogno di maggiore spazio.

1.4.3 Data warehouse

I dati provenienti dalle differenti sorgenti, una volta lavorati tramite i processi e gli strumenti ETL, vengono archiviati in uno specifico database, distinto e separato da quelli di carattere operativo come possono essere i transazionali o i relazionali. Tale collezione di dati, o repository, prende il nome di "data warehouse"e consente agli knowledge workers (come analisti, gestori, amministratori e dirigenti) di trovarvi successivamente ausilio, in ambito decisionale e strategico [92], tramite diverse operazioni quali l'esecuzione di query, l'effettuazione di analisi e la generazione di report. Il data warehouse può pertanto risultare utile in differenti contesti come nel commercio, nella manifattura, nei servizi finanziari, trasporti, telecomunicazione, etc [61]. Al fine di essere

utilizzato con tale intento, il data warehouse deve disporre di precise proprietà al termine della sua progettazione. La divisione dai depositi operativi è ad esempio dovuta per permettere di incrementarne le prestazioni e migliorarne la gestione; in questa visione di sostegno, infatti, si hanno ricerche più complesse e ulteriori tecniche di accesso, l'efficienza della manipolazione dei dati è inoltre maggiore data la possibilità di colmare le informazioni mancanti e di migliorarne, al contempo, la qualità stessa oltre che il loro consolidamento [88]. Altri aspetti rilevanti della data warehouse sono [61, 67]:

- **Orientatamento ai soggetti di interesse:** non ci si focalizza sui processi aziendali ma sui clienti di destinazione, sui prodotti finiti o sui fornitori a seconda della ragione dell'organizzazione.
- **Integrata:** le sorgenti rappresentanti le fonti possono essere particolarmente numerosi e di caratteristiche pertanto differenti. È conseguentemente necessario che i loro dati vengano rappresentati in modo univoco.
- **Dati aggregati**
- **Correlata alla variabile del tempo:** l'aggiornamento dei dati avviene periodicamente.
- **Non volatile:** non sono realizzate operazioni di modifica o aggiunta, ci si vincola alla sola lettura.

La progettazione invece del data warehouse, capace successivamente di poter perseguire le caratteristiche appena citate, può avvenire per mezzo di due distinti approcci [61], a seconda dei contesti da voler analizzare e, inevitabilmente, dalle risorse a disposizione.

- **Top-Down:** è il metodo che consente di risalire ad un'analisi di tipo globale dell'organizzazione in considerazione. I costi risultano essere elevati in relazione anche delle tempistiche mentre la progettazione è complessa.

- **Bottom-Up:** più semplice nella fase di realizzazione rispetto l'alternativa, permette di focalizzarsi su precisi settori e non sull'intero complesso. I costi sono minori ed il data warehouse viene completato in modo incrementale.

Analogamente alla progettazione, l'architettura dovrà essere tale da favorire la capacità di elaborazione; precise caratteristiche dovranno pertanto essere presenti. La prima è la già citata **separazione** tra l'elaborazione analitica e quella transazionale alla quale segue la **scalabilità** dei sistemi hardware e software in relazione del volume dei dati da dover gestire. Si ha inoltre il bisogno che il data warehouse risponda alla necessità di **estendibilità** alle diverse tipologie di applicazioni, senza dover ricorrere ad una nuova progettazione. Per ultimo, ma non per importanza, è la **sicurezza**, rilevante nel controllo degli accessi [67, 89]. Le architetture sono di tre tipologie:

- **Architettura a 1 livello,** è maggiormente indicato se l'intento principale è la riduzione della ridondanza e quindi del volume di dati. Il data warehouse risulta qui essere virtuale o meglio è realizzato come se fosse una vista multidimensionale generato da un middleware: uno strato di elaborazione intermedio. Problema cardine è che non dispone della separazione tra l'elaborazione analitica OLAP e transazionale OLTP [91].

- **Architettura a 2 livelli,** è prettamente utilizzato al fine di evidenziare e distanziare due precise fasi rispettivamente relative alle sorgenti e al data warehouse. Ciò è possibile per mezzo di un'ulteriore fase, intermedia, che riguarda le operazioni ETL. A concludere il processo, vi è una quarta e ultima fase attinente all'analisi dei dati estrapolati, elaborati ed archiviati [61].

  La terza fase può disporre di diversi supporti quali la presenza dei **data mart**, costituiti da parte dei dati provenienti dal data warehouse stesso (dedicati a precisi ambiti dell'organizzazione) oppure da **meta dati**, strumenti che forniscono informazioni sulle sorgenti come ad esempio i meccanismi di accessibilità.

  Si hanno diverse varianti di questa architettura:

- **Data mart indipendenti,** questa consiste nell'alimentare i data mart direttamente dalle sorgenti. Le loro dimensioni ridotte infatti incrementano notevolmente le prestazioni rispetto all'utilizzo del solo data warehouse primario, oltre che snellirne le fasi progettuali [89, 92].
- **Data mart dipendenti,** al contrario qui i data mart sono correlati direttamente al data warehouse e non alle sorgenti. Permettono di delineare inoltre le informazioni per particolari utenti e quindi interrogazioni [89, 92].

- **Architettura a 3 livelli,** presenta un ulteriore strato tra il primo livello, rappresentato dalle sorgenti, ed il terzo ovvero il data warehouse. A valle delle operazioni ETL infatti, vi è l'operational data store o dei dati riconciliati che alimenta il data warehouse. La separazione è più netta ma la ridondanza è maggiore [61].

1.4.4 Progettazione concettuale

I dati delle differenti sorgenti al fine che siano il più possibile uniformi, strutturati e pertanto aggregati all'interno del data warehouse, seguono una precisa composizione dettata dalla corrispondente progettazione concettuale di carattere multidimensionale. Il delineamento e l'organizzazione dei dati consente infatti sia una maggiore agevolazione nella comprensione e gestione dell'informazione da parte degli knowledge worker, che andranno successivamente ad usufruirne, sia l'incremento dell'efficienza e pertanto delle prestazioni dell'intero sistema.

Il modello maggiormente adoperato in tale contesto risulta essere il Dimensional Fact Model (DFM) realizzato da Rizzi e Golfarelli nel 1998 [61]. Il modello DFM si presta maggiormente alla progettazione concettuale del data warehouse e degli eventuali data mart a differenza, invece, di quanto possa offrire il modello Entity Relationship (ER) che, seppur notevolmente utilizzato per il delineamento delle strutture delle basi di dati, non è spesso capace di rappresentare la totalità delle relazioni fra le diverse entità [61]. Il modello DFM fa quindi riferimento all'analisi multidimensionale la quale può essere

interpretata mediante cubi n-dimensionali i cui concetti costituenti risultano essere principalmente tre:

- **Fatto:** rappresenta un gruppo di eventi sul quale incentrare le analisi e corrisponderebbe all'intero cubo n-dimensionale.
- **Misura**: proprietà atomica tale pertanto da quantificare il rispettivo fatto e si riscontra nelle celle all'interno del cubo.
- **Dimensione**: rappresenta la proiezione del fatto verso aspetti di interesse ovvero ne descrive le coordinate di analisi.

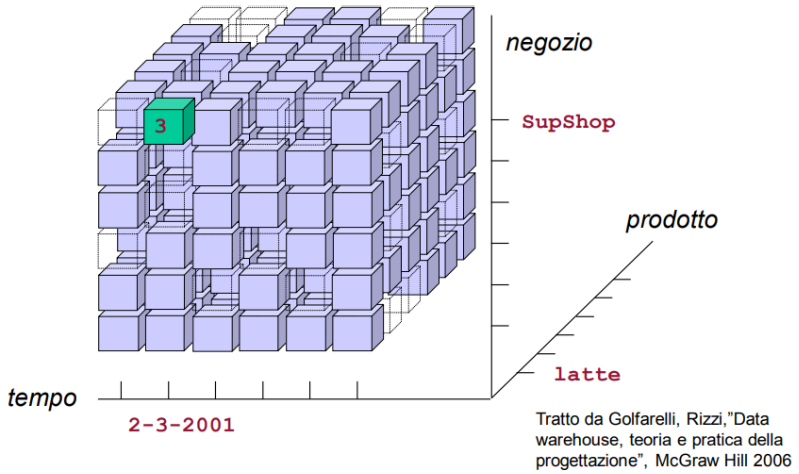


Figura 5.

L'implementazione della Dimensional Fact permette il passaggio ad una rappresentazione più compatta dove si constata maggiormente un ulteriore concetto: la **gerarchia.** Quest'ultimo altro non corrisponderebbe che all'aggregazione di un sottoinsieme di dimensioni [61, 76].

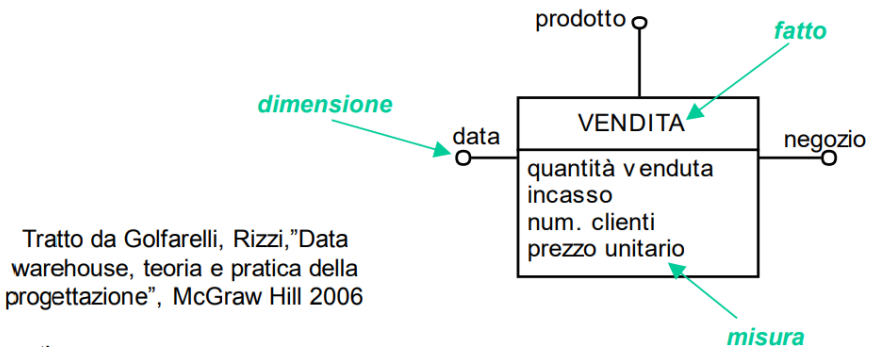
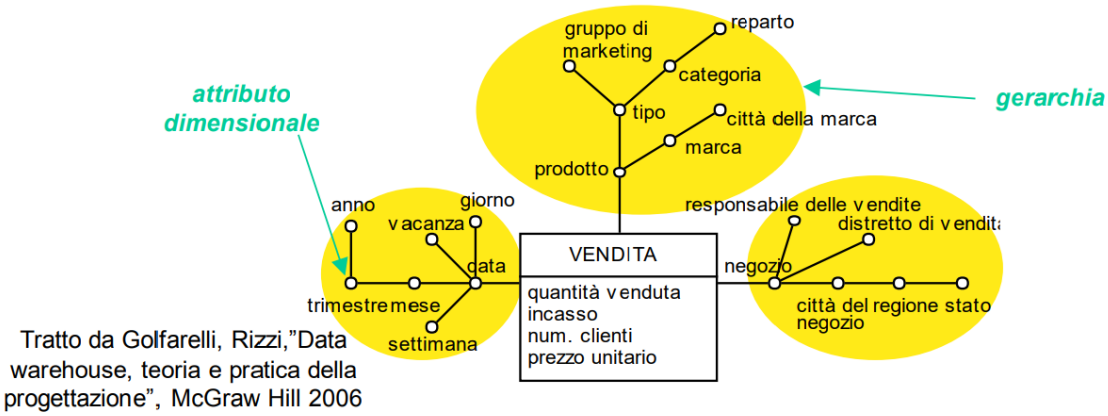
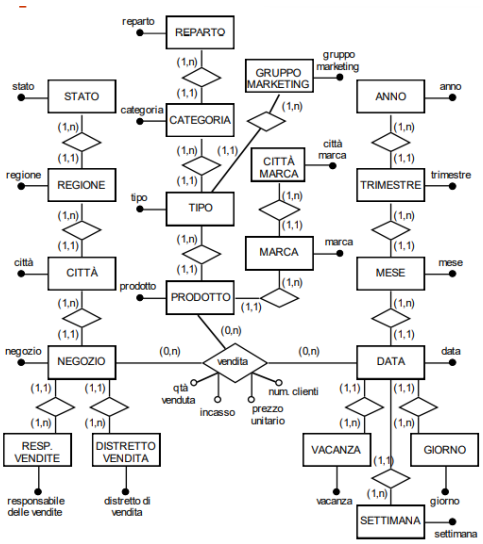


Figura 6.



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Figura 7.



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Figura 8.

1.4.5 Progettazione logica

Una volta descritte le caratteristiche del mondo reale, come visto nella progettazione concettuale, ci si proietta verso la definizione della struttura dei dati dei DBMS mediante la progettazione logica [61]. Questa fase risulta difatti estremamente delicata in quanto un'eventuale errata costruzione dell'intelaiatura che lega i dati, e quindi della modalità di archiviazione, può comportare inefficienze per le fasi di analisi successive (le esecuzioni di query e le opzioni di elaborazione rientrano ad esempio tra queste). Del resto, la mole di dati necessari per la comprensione e la direzione delle scelte

organizzative, nel contesto come detto del data warehouse, comporta adeguati strumenti e tecniche che trovano riscontro nei sistemi di tipo OLAP (On-Line Analytical Processing) [92]. Le funzioni di supporto da questi fornite infatti, comprendono un elevato numero di operazioni, rappresentate da complesse query da eseguire su grandi volumi di dati; tali tecniche software, utilizzando una visione dimensionale dei dati aggregati, consentono pertanto un rapido accesso alle diverse informazioni maggiormente di interesse. Nei sistemi OLAP si distinguono due principali soluzioni, a seconda delle tecnologie adoperate, alle quali se ne aggiunge una terza, intermedia alle precedenti [61, 67, 89, 91, 92, 93, 94].

- **ROLAP (Relational Online Analytical Processing):** si basano sulle strutture dei modelli relazionali tradizionali. I concetti multidimensionali sono rappresentati mediante tabelle e pertanto bidimensionalmente. I dati provengono direttamente dal data warehouse principale mentre viene dinamicamente realizzata la vista di tipo multidimensionale. Risulta più flessibile ma presenta un sovraccarico computazionale maggiore a causa della gestione delle numerose tabelle. È maggiormente efficiente, in termini di spazio per un database di grandi dimensioni, se i suoi aggregati hanno una distribuzione concetrata nel cubo di dati. Poiché sono un'estensione del database relazione, è anche possibile in questo adoperare il linguaggio query standard (SQL).

  I dati dei datamart sono qui organizzati mediante uno schema dimensionale riconducibile ad uno **schema a stella**. Questa corrisponde ad una rappresentazione di una tabella principale, denominata "tabella dei fatti", che memorizza i fatti misurabili (numerici, additivi e continui) di un processo. Le misure sono gli attributi delle tabella e la chiave primaria è la combinazione delle chiavi esterne delle tabelle "dimensioni". Queste ultime riguardano i dati riguardanti le dimensioni delle analisi e possono essere descrittivi, testuali o discreti. Si ha una tabella per ogni dimensione, ogni chiave primaria è artificiale mentre la rappresentazione è denormalizzata, ovvero sono presenti ridondanze tra i dati. Al fine di ridurre la ripetizione di informazioni, si ricorre ad un ulteriore

schema, considerato un'estensione del precedente, ovvero lo **schema a fiocco di neve (snowflake)**. La sua peculiarità corrisponderebbe ad una rappresentazione più normalizzata delle tabelle "dimensione" permettendo di raggiungere, dalla "tabella dei fatti", tutte le tabelle delle dimensioni. Solitamente la "normalizzazione" è sconsigliata in quanto il costo di eseguire più join può risultare significativo e vi è comunque una riduzione dello spazio. Può d'altro canto essere però utile quando porzioni di una gerarchia sono condivise tra più dimensioni.

- **MOLAP (Multidimensional Online Analytical Processing):** in questi sistemi l'allocazione avviene tramite vettore multidimensionale con accesso posizionale. I dati vengono memorizzati e gestiti dagli MDDB del database proprietario ovvero da database multidimensionali che sono associabili a dei cubi o in alcuni casi ipercubi, a seconda delle dimensioni del modello rappresentato. La vista multidimensionale non è realizzata dinamicamente ma risulta già presente e pertanto statica. Le prestazioni risultano essere più efficienti e la velocità di accesso maggiore in quanto non sono necessarie complesse query ma sono sufficienti le specifiche coordinate per risalire alla combinazione di interesse in modo diretto. Problematica non poco rilevante è data dallo spreco di memoria dipendente dalla cosidetta "sparsità di dati", terminologia con la quale s'intende la percentuale di celle che contengono valore nullo, dato che corrispondono ad eventi mai accaduti (per esempio un cubo o ipercubo con sparsità del 25% sta a significare che un quarto delle sue celle non conterranno valori). La difficoltà della sparsità dei dati può essere in parte risolta suddividendo il cubo n-dimensionale in sottocubi n-dimensionali denominati chunk. Ulteriore problematica è l'assenza di un'interfaccia come SQL oltre che l'assenza di join e limitazioni di scalabilità.
- **HOLAP (Hybrid OLAP):** risultano questi essere una combinazione dei due differenti approcci di archivazione, ovvero ROLAP e MOLAP. Si memorizza mediante database relazionali e database multidimensionali.

I diversi server adoperano pertanto tecniche differenti per la salvaguardia e la gestione dell'informazione ma le operazioni sui dati rimangano pressochè le medesime. Queste difatti si riversano sulle rappresentazioni dei cuboidi.

- **Roll-up:** aggregazione dei dati per mezzo della riduzione dei loro dettagli mediante ad esempio l'eliminazine di una delle dimensioni presenti o l'applicazione di una serie di formule.
- **Drill-down:** aumento dei dettagli per mezzo dell'introduzione di una nuova dimensione o di un suo attributo con la conseguente riduzione della gerarchia. È riconducibile al comando "GROUPBY" del linguaggio SQL.
- **Slicing:** selezione e proiezione di una sola dimensione del cuboide, andando quindi a realizzare un sottocubo dell'iniziale. Si riduce così il volume dei dati.
- **Dicing:** similmente allo slicing, si differenzia invece per la selezione di più dimensioni del medesimo cuboide come pertanto una combinazione di predicate.
- **Pivoting:** modifica dell'orientamento del cuboide, della struttura multidimensionale, secondo una specifica dimensione ma senza intaccare il livello di dettaglio. Si ha in questo modo una visualizzazione più chiara.

1.4.6 Data mining

Quanto visto finora consente di affermare che la tecnologia del data warehouse include diverse operazioni: dalla pulizia all'integrazione dei dati, fino all'elaborazione analitica (OLAP), ovvero tecniche di analisi con funzionalità di riepilogo, consolidamento e aggregazione, nonché la capacità di visualizzare le informazioni da diverse angolazioni. Sebbene gli strumenti OLAP supportino l'analisi multidimensionale e il processo decisionale, sono necessari strumenti di analisi dei dati aggiuntivi per un'analisi approfondita, che forniscano la classificazione dei dati, il clustering, il rilevamento di valori anomali ed i cambiamenti nei dati nel tempo [61, 67, 89, 91, 92, 93, 94]. La rapida acquisizione di una grande quantità di dati infatti [102], definiti da più variabili (dati multivariati o megavariati), e la sua corretta preelaborazione e trattamento, hanno

difatto accelerato l'origine a una nuova disciplina scientifica nota come "scienza analitica dei big data", che ha avuto un forte impatto in molti ambiti. Diversi nomi sono stati convenzionalmente applicati per riferirsi all'elaborazione dei dati multivariata. Tuttavia, negli ultimi anni, i termini "data mining" e "machine learning" [95] hanno prevalso e vengono utilizzati più frequentemente. Il termine data mining, sul quale ci concentreremo con stretto riferimento al libro di Jiawei Han, Micheline Kamber, Jian Pei ovvero "Data Mining concepts and techniques" della Morgan Kaufmann e pubblicata da Elsevier [94], si riferisce all'estrazione di conoscenze significative da informazioni utili ma non evidenti, che sono nascoste all'interno di grandi set di dati. Il termine di machine learning è noto come i metodi computazionali coinvolti nella gestione di dati vasti nel modo più intelligente (sviluppando algoritmi) per ricavare intuizioni fruibili; non solo, è inoltre spesso considerato il termine data mining come sinonimo di un altro termine comunemente usato ovvero "knowledge discovery dai dati" o KDD, mentre altri considerano il data mining semplicemente come un passaggio essenziale nel processo di knowledge discovery. Il processo di knowledge discovery è mostrato come una sequenza iterativa dei seguenti passaggi:

1. Data cleaning (per rimuovere rumore e dati incoerenti).
2. Data integration (dove possono essere combinate più fonti di dati).
3. Data selection (dove i dati rilevanti per l'attività di analisi vengono recuperati dal file Banca dati).
4. Data transformation (dove i dati vengono trasformati e consolidati in forme idonee per il mining eseguendo operazioni di riepilogo o aggregazione).
5. Data mining (un processo essenziale in cui vengono applicati metodi intelligenti per estrarre modelli di dati).
6. Valutazione dei modelli (per identificare i modelli veramente interessanti che rappresentano la conoscenza in base alle misure di interesse).
7. Presentazione della conoscenza (in cui vengono utilizzate tecniche di visualizzazione e rappresentazione della conoscenza per presentare la conoscenza estratta agli utenti).

I passaggi da 1 a 4 sono diverse forme di pre-elaborazione in cui i dati vengono preparati per il mining. La fase di data mining, in ogni caso, può interagire con l'utente o con una knowledge base. I modelli interessanti vengono presentati all'utente e possono rappresentare l'effettiva conoscenza. D'altro canto però, il data mining è presentato come un singolo passo nel processo di scoperta della conoscenza in quanto essenziale perché scopre modelli nascosti per la valutazione. Tuttavia, nell'industria, nei media e nell'ambiente di ricerca, il termine data mining viene spesso utilizzato per riferirsi all'intero processo di scoperta della conoscenza appena descritto.

Come tecnologia generale invece, la data mining può essere applicata a qualsiasi tipo di dati purché questi siano significativi per un'applicazione di destinazione. Le forme di dati più elementari per il mining appartengo a database di vario tipo o a data warehouse. Il data mining può essere applicato però anche ad altre forme di dati (ad es. flussi di dati, dati in sequenza, dati grafici o in rete, dati spaziali, dati di testo, dati multimediali e domini WWW). Le funzionalità della data mining vengono utilizzate per specificare i tipi di pattern da trovare nelle attività di data mining. In generale, tali attività possono essere classificate in due categorie: descrittive e predittive. Le attività di mining descrittive caratterizzano le proprietà dei dati in un data set di destinazione. Le attività di mining predittivo eseguono l'induzione sui dati correnti per fare previsioni.

I pattern frequenti, come suggerisce il nome, sono modelli che si verificano frequentemente nei dati. Esistono molti tipi di pattern frequenti, inclusi set di elementi, sottosequenze (note anche come pattern sequenziali) e sottostrutture. Un set di articoli frequente si riferisce in genere a un insieme di articoli che spesso compaiono insieme in un set di dati transazionali, ad esempio latte e pane, che vengono spesso acquistati insieme nei negozi di alimentari da molti clienti. Una sottosequenza che si verifica frequentemente invece, può essere rappresentata dal modello secondo cui i clienti tendono ad acquistare prima un laptop, seguito da una fotocamera digitale e quindi una scheda di memoria. Una sottostruttura può fare riferimento a diverse forme (ad esempio grafici, alberi o reticoli) che possono essere combinate con set di elementi o sottosequenze. Se una sottostruttura si verifica frequentemente, viene chiamata pattern

strutturato. L'estrazione di modelli frequenti porta alla scoperta di associazioni e correlazioni interessanti all'interno dei dati.

I principali pattern possono essere associati principalmente a tre tipologie:

- **Regole associative:** la loro realizzazione consiste nel trovare innanzitutto degli elementi, dagli insiemi chiamati transazioni, con alta frequenza come potrebbero essere A e B (che soddisfano una soglia minima di apparizioni o una percentuale delle tuple rilevanti per l'attività), da cui vengono generate delle correlazioni forti sotto forma del tipo di:

  $$A \Rightarrow B$$

  Queste regole soddisfano anche una soglia minima di affidabilità (un valore predefinito probabilità di soddisfare B a condizione che A sia soddisfatto) e altro non consentono, in conclusione, che l'individuazione di gruppi di affinità tra gli oggetti stessi di una transazione. Alla presenza dell'elemento A pertanto, si suppone che sia presente nella medesima attività anche l'elemento B. Si distinguono due particolari indicatori per le regole di associazione:

  - Supporto: corrisponderebbe alla frequenza degli elementi A e B.
  - Confidenza: relativa alla percentuale di transazioni, con all'interno l'elemento A, che dispone anche dell'elemeno B.

  Un terzo indicatore è il "lift" che permette di conoscere indicativamente il valore di indipendenza tra gli elementi in considerazione.

  Le associazioni possono essere utilizzate in diversi e vari contesti quali il market basket analysis, cross-marketing, progettazione di layout e cataloghi e vengono adoperate al fine di analizzare gli antecedenti (quale elemento comporta B), i conseguenti (quale elemento corrisponde ad una transazione con all'interno l'elemento A) o per determinare una visione di alto livello delle correlazioni presenti [97].

- **Classificazione:** è una forma di analisi dei dati che estrae modelli che descrivono classi, ad ognuna della quali associare un set di dati. Tali modelli, chiamati

classificatori, prevedono etichette di classe di tipo categoriali (discrete, non ordinate). Ad esempio, si può creare un modello di classificazione per distinguere le richieste di prestito bancario come sicure o rischiose. Tale analisi può aiutare a fornire una migliore comprensione dei dati in generale. La classificazione ha numerose applicazioni, tra cui il rilevamento di frodi, marketing mirato, previsione delle prestazioni, produzione e diagnosi medica.

Le categorie determinate per la distinzione dei dati, possono essere rappresentate da valori discreti come detto, dove l'ordinamento tra i valori non ha significato. Ad esempio, i valori 1, 2 e 3 utilizzati per rappresentare i trattamenti A, B e C nel caso di una soluzione per una determinata problematica medica, non corrispondono ad alcun ordinamento implicito.

Molte volte, inoltre, la classificazione viene associata concettualmente alla previsione numerica. Supponendo per esempio che un responsabile marketing voglia prevedere quanto spenderà un determinato cliente durante una vendita presso la propria azienda, il manager si accingerà ad intraprende un'attività di analisi dei dati al fine di determinare una previsione numerica, in cui il modello costruito andrà a realizzare una funzione a valori continui o un valore ordinato, al contrario di un'etichetta di classe, come visto precedentemente con i diversi trattamenti. Questo modello è pertanto un predittore. Nella previsione numerica è più frequentemente adoperata l'analisi di regressione come metodologia statistica a differenza di quanto accade invece nella tecnica di classificazione.

La classificazione dei dati è un processo suddiviso in due fasi: la prima di apprendimento, in cui viene costruito un modello per la distinzione (ovvero il classificatore) che descrive un insieme predeterminato di classi o concetti di dati, e una seconda di classificazione, in cui il modello viene utilizzato per prevedere le etichette di classe per i diversi dati.

Nella prima fase si ha modo di realizzare il modello mediante diverse tecniche come l'albero decisionale, le regole di associazione, reti neurali, support vector machines (SVM) e probabilità. Necessario è però il training test nel quale si verificano le condizioni e l'affidabilità del modello stesso prima di poter passare alla seconda fase che permette l'applicazione del modello su set di dati di cui è ignota la classe di appartenenza.

- **Albero decisionale:** è una struttura ad albero simile ad un diagramma di flusso, in cui ogni nodo interno (nodo non foglia) denota un test su un attributo, ogni ramo rappresenta un risultato del test e ogni nodo foglia (o nodo terminale) contiene un'etichetta di classe. Il nodo più in alto in un albero è il nodo radice. I nodi interni sono indicati da rettangoli e i nodi foglia sono indicati da ovali. Alcuni algoritmi dell'albero decisionale producono solo alberi binari (dove ogni nodo interno si dirama esattamente verso altri due nodi), mentre altri possono produrre alberi non binari. Data una tupla, X, per la quale l'etichetta della classe associata è sconosciuta, i valori degli attributi della tupla vengono testati rispetto all'albero decisionale. Viene tracciato un percorso dalla radice a un nodo foglia, che contiene la previsione della classe per quella tupla. Gli alberi decisionali possono essere facilmente convertiti in regole di classificazione. La costruzione di classificatori dell'albero decisionale non richiede alcuna conoscenza del dominio o impostazione di parametri, ed è quindi appropriata per la scoperta esplorativa della conoscenza. Gli alberi decisionali possono gestire dati multidimensionali. La loro rappresentazione della conoscenza acquisita sotto forma di albero è intuitiva e generalmente facile da assimilare. Le fasi di apprendimento e classificazione dell'induzione dell'albero decisionale sono semplici e veloci. In generale, i classificatori dell'albero decisionale hanno una buona precisione. Tuttavia, il successo dell'utilizzo può dipendere dai dati disponibili. Gli algoritmi di induzione dell'albero decisionale sono stati

utilizzati per la classificazione in molte aree di applicazione come medicina, produzione, analisi finanziaria, astronomia e biologia molecolare. Gli alberi decisionali sono la base di diversi sistemi di regole commerciali. Quando vengono creati alberi decisionali, molti rami possono riflettere rumore o valori anomali nei dati di addestramento. La potatura degli alberi tenta di identificare e rimuovere tali rami, con l'obiettivo di migliorare l'accuratezza della classificazione su dati.

- **Classificatori bayesiani:** sono classificatori di tipo statistico. Possono prevedere infatti la probabilità che una data tupla appartenga a una particolare classe. La classificazione bayesiana si basa sul teorema di Bayes i cui algoritmi di classificazione rendono il classificatore bayesiano comparabile in termini di prestazioni con l'albero decisionale e classificatori di reti neurali. I classificatori bayesiani hanno anche mostrato alta precisione e velocità quando applicati a database di grandi dimensioni; inoltre presumono che l'effetto di un valore di attributo su una data classe sia indipendente dai valori degli altri attributi. Questa ipotesi è chiamata indipendenza condizionata dalla classe ed è fatto per semplificare i calcoli coinvolti.
- **Reti neurali:** originariamente ideato da psicologi e neurobiologi che hanno cercato di sviluppare e testare analoghi computazionali dei neuroni, è un insieme di unità di input / output collegate in cui ogni connessione ha un peso associato ad essa. Durante la fase di apprendimento, la rete calibra i pesi in modo da poter prevedere la corretta etichetta di classe delle tuple di input.

  Le reti neurali richiedono lunghi tempi di calibratura e sono quindi più adatte ad applicazioni dove ciò è fattibile. Richiedono una serie di parametri che in genere sono meglio determinati empiricamente. Le reti neurali sono state criticate per la loro scarsa interpretabilità ad esempio per la difficoltà di interpretazione del significato simbolico dietro i pesi.

Queste caratteristiche inizialmente rendevano le reti neurali meno desiderabili per il data mining.

I vantaggi delle reti neurali, tuttavia, includono la loro elevata tolleranza ai dati rumorosi e la loro capacità di classificare modelli su cui non sono stati calibrati. Possono essere usati quando si ha poca conoscenza delle relazioni tra attributi e classi. Sono adatti per input e output a valori continui, a differenza della maggior parte degli algoritmi dell'albero decisionale. Hanno avuto successo su una vasta gamma di dati del mondo reale, tra cui il riconoscimento dei caratteri scritti a mano, la patologia e la medicina di laboratorio. Gli algoritmi di rete neurale sono intrinsecamente paralleli; le tecniche di parallelizzazione possono essere utilizzate per accelerare il processo di calcolo.

Questi fattori contribuiscono all'utilità delle reti neurali per la classificazione e la previsione numerica nel data mining. Esistono molti tipi diversi di reti neurali e algoritmi di rete neurale. L'algoritmo di rete neurale più popolare è la backpropagation, che ha guadagnato reputazione negli anni '80.

- **Clustering:** l'analisi dei cluster è il processo di partizionamento di un insieme di oggetti dati (o osservazioni) in sottoinsiemi. Ogni sottoinsieme è un cluster, in modo tale che gli oggetti in un cluster sono simili tra loro, ma dissimili dagli oggetti di altri cluster. L'insieme di cluster risultante da un'analisi viene denominato clustering. In questo contesto, diversi metodi di clustering possono generare diversi risultati sullo stesso set di dati, ognugno dei quali seguendo algoritmi differenti per il partizionamento. Il clustering è utile in quanto può portare alla scoperta di gruppi precedentemente sconosciuti all'interno dei dati. L'analisi dei cluster è stata inoltre ampiamente utilizzata in molte applicazioni come la business intelligence, appunto, per il riconoscimento di pattern di immagini, ricerca sul Web, biologia e sicurezza. Considerata ad esempio una società di consulenza con un gran numero di progetti, per migliorarne la gestione il processo di clustering può essere applicato per suddividere i progetti in categorie basate

sulla somiglianza, in modo che l'audit e la diagnosi del progetto (per migliorarne la consegna e i risultati) possano essere condotti in modo efficace.

Come funzione di data mining, l'analisi dei cluster può essere utilizzata come strumento autonomo per ottenere informazioni sulla distribuzione dei dati, osservare le caratteristiche di ciascun cluster e concentrarsi su un particolare gruppo per ulteriori analisi. In alternativa può servire come fase di pre-elaborazione per altri algoritmi, come la caratterizzazione, la selezione di sottoinsiemi di attributi e la classificazione, che poi opererebbero sui cluster rilevati e sugli attributi o caratteristiche selezionati.

Il processo di clustering in alcune applicazioni è anche chiamato segmentazione dei dati perché suddivide grandi set di dati in gruppi in base alla loro somiglianza. Può anche essere utilizzato per il rilevamento dei valori anomali, laddove i valori anomali (valori "lontani" da qualsiasi cluster) possono essere più interessanti dei casi comuni. Le applicazioni del rilevamento dei valori anomali vengono abitualmente utilizzati per il rilevamento di frodi con carte di credito e il monitoraggio di attività criminali nel commercio elettronico.

A causa dell'enorme quantità di dati raccolti nei database, l'analisi dei cluster è diventata recentemente un argomento molto attivo nella ricerca sul data mining. Strumenti di analisi dei cluster basati su k-means, k-medoids e molti altri metodi sono stati incorporati anche in molti pacchetti software o sistemi di analisi statistica, come S-Plus, SPSS e SAS. Nel data mining, gli sforzi si sono concentrati sulla ricerca di metodi per un'analisi cluster efficiente ed efficace in database di grandi dimensioni.

Esistono molti algoritmi di clustering in letteratura che possono però sovrapporsi in modo che un metodo possa avere caratteristiche di diverse categorie. Tuttavia, è utile presentare un quadro relativamente organizzato dei metodi di raggruppamento. In generale, i principali metodi fondamentali di raggruppamento possono essere classificati nelle seguenti categorie:

- **Metodi di partizionamento:** dato un insieme di n oggetti, un metodo di partizionamento costruisce k partizioni dei dati, dove ogni partizione

rappresenta un cluster e k ≤ n. Cioè, divide i dati in k gruppi in modo tale che ogni gruppo deve contenere almeno un oggetto. In altre parole, i metodi di partizionamento conducono il partizionamento a un livello sui set di dati.

I metodi di partizionamento di base adottano in genere la separazione esclusiva dei cluster dove ogni oggetto deve appartenere esattamente a un gruppo. Questo requisito può essere attenuato, ad esempio, nelle tecniche di partizionamento fuzzy.

La maggior parte dei metodi di partizionamento sono basati sulla distanza. Dato k, il numero di partizioni da costruire, un metodo di partizionamento crea un partizionamento iniziale. Quindi utilizza una tecnica di rilocazione iterativa che tenta di migliorare il partizionamento spostando gli oggetti da un gruppo a un altro. Il criterio generale di un buon partizionamento è che gli oggetti nello stesso cluster sono "vicini" o correlati tra loro, mentre gli oggetti in diversi cluster sono "lontani" o molto diversi. Il raggiungimento dell'ottimalità globale nel clustering basato sul partizionamento è spesso proibitivo dal punto di vista computazionale, e richiede potenzialmente un'enumerazione esaustiva di tutte le possibili partizioni. Invece, la maggior parte delle applicazioni adotta metodi popolari, come gli algoritmi k-means e k-medoids, che migliorano progressivamente la qualità del clustering e si avvicinano a un ottimo locale. Questi metodi di clustering funzionano bene per trovare cluster in database di piccole e medie dimensioni. Per trovare cluster con forme complesse e per set di dati molto grandi, i metodi basati sul partizionamento devono essere estesi.

- **Metodi gerarchici:** qui si crea una scomposizione gerarchica del dato. Un metodo gerarchico può essere classificato come agglomerativo o divisivo, in base a come si forma la scomposizione gerarchica.
  - **Approccio agglomerativo:** inizia con ciascun oggetto che forma un gruppo separato. Successivamente si uniscono gli oggetti o i gruppi vicini l'uno all'altro, fino a quando tutti i gruppi vengono uniti in uno

(il livello più alto della gerarchia) o una condizione di terminazione è valida.

- **Approccio divisivo:** chiamato anche approccio top-down, inizia con tutti gli oggetti nello stesso cluster. In ogni iterazione successiva, un cluster viene suddiviso in cluster più piccoli, fino a quando alla fine ogni oggetto si trova in un cluster, o una condizione di terminazione si mantiene.

I metodi di raggruppamento gerarchico possono essere basati sulla distanza o sulla densità e sulla continuità. Varie estensioni di metodi gerarchici considerano anche il raggruppamento in sottospazi. I metodi gerarchici soffrono del fatto che una volta eseguito un passaggio (unione o divisione), non può essere annullato.

- **Metodi basati sulla densità:** la maggior parte dei metodi di partizionamento raggruppa gli oggetti in base alla distanza tra gli oggetti mentre altri metodi di raggruppamento sono stati sviluppati basandosi sulla nozione di densità. La loro idea generale è di continuare a far crescere un dato cluster fintanto che la densità (numero di oggetti o punti dati) nel "vicinato" supera una certa soglia. Ad esempio, per ogni punto dati all'interno di un dato cluster, l'intorno di un dato raggio deve contenere almeno un numero minimo di punti. Un tale metodo può essere utilizzato per filtrare il rumore o valori anomali.
- **Metodi basati su griglia:** i metodi basati su griglia quantizzano lo spazio dell'oggetto in un numero finito di celle che formano una struttura a griglia. Il vantaggio principale di questo approccio è il suo tempo di elaborazione veloce, che è tipicamente indipendente dal numero di oggetti dati e dipende solo dal numero di celle in ciascuna dimensione nello spazio quantizzato.

  L'utilizzo delle griglie è spesso un approccio efficiente a molti problemi di data mining spaziale, incluso il clustering. Pertanto, i metodi basati su

griglia possono essere integrati con altri metodi di raggruppamento come metodi basati sulla densità e metodi gerarchici.

Alcuni algoritmi di clustering integrano le idee di diversi metodi di clustering, così che a volte è difficile classificare un dato algoritmo come appartenente in modo univoco a una sola categoria di metodi di clustering. Inoltre, alcune applicazioni possono avere criteri di clustering che richiedono l'integrazione di diverse tecniche di clustering.

- **Serie storiche:** permettono l'individuazione di pattern ricorrenti o anormali in sequenze di dati elaborate e complesse. L'applicazione può riguardare numerevoli ambiti quali la rilevazione di problematiche all'interno di un sistema di monitoraggio, l'individuazione di eventuali correlazioni tra periodi temporali differenti o di schemi associativi tra titoli di borsa fino a giungere allo studio sui percorsi di navigazione riguardanti un sito web.

1.4.7 Analisi what-if

Un numero crescente di organizzazioni sente la necessità di ottenere informazioni rilevanti sul proprio business futuro, finalizzate a pianificare strategie ottimali per raggiungere i propri obiettivi. In particolare, per poter valutare in anticipo l'impatto di una mossa strategica o tattica, i decisori necessitano di sistemi previsionali affidabili. I data warehouse (DW), che in effetti hanno svolto un ruolo di primo piano all'interno delle piattaforme di business intelligence nel sostenere il processo decisionale, hanno lo scopo di supportare un'analisi dettagliata dei dati passati, quindi non sono in grado di fornire anticipazioni di tendenze future [100]. E qui che entra in gioco l'analisi what-if.

In poche parole, l'analisi what-if può essere descritta come una simulazione di dati il cui obiettivo è quello di ispezionare il comportamento di un sistema complesso (l'attività aziendale o una parte di esso) sotto alcune ipotesi date (chiamate scenari). Più pragmaticamente, l'analisi what-if misura come i cambiamenti delle variabili indipendenti influiscono su un insieme di variabili dipendenti con riferimento a un dato modello di simulazione; tale modello è una rappresentazione semplificata del business,

relazionata sulla base dei dati storici dell'impresa. L'analisi what-if non deve essere confusa con l'analisi della sensibilità, volta a valutare quanto sia sensibile il comportamento del sistema ad una piccola variazione di uno o più parametri. Un semplice esempio di query what-if, nel dominio in questo caso del marketing, può essere il come cambierebbero i profitti se si effettuasse una promozione 3 × 2 per una settimana, su alcuni prodotti in vendita [99]. Inoltre, vi è un'importante differenza tra analisi what-if e semplice previsione, ampiamente utilizzata soprattutto in ambito bancario e assicurativo. Infatti, mentre la previsione è normalmente eseguita estrapolando tendenze dalle serie storiche memorizzate nei sistemi informativi, l'analisi what-if richiede di simulare fenomeni complessi i cui effetti non possono essere determinati semplicemente come una proiezione di dati passati, che a sua volta richiede di costruire un modello di simulazione in grado di riprodurre, con un'approssimazione soddisfacente, il comportamento reale dell'organizzazione. Per lo stesso motivo, la progettazione di applicazioni what-if è anche più complessa di quella dei DW convenzionali, che si basa solo su un modello statico di business. Sebbene non siano disponibili in commercio piattaforme what-if dedicate, alcuni strumenti di data warehousing o di previsione sono stati estesi con le caratteristiche dell'analisi what-if come SAP BPS, SAS Forecast Server o ancora QlikView. Uno strumento per l'analisi what-if dovrebbe avere almeno le seguenti caratteristiche [100]:

- Supportare una serie di tecniche per esprimere e costruire modelli di simulazione, oltre a un linguaggio per estendere ulteriormente le capacità di modellazione.
- Supportare i decisori nella formulazione di scenari ipotetici sul modello.
- Supportare l'aggiornamento interattivo dei dati.
- Consentire ai decisori di aggregare e disaggregare gerarchicamente le previsioni e vedere l'impatto delle modifiche a ogni livello.
- Supportare tecniche statistiche per valutare quanto siano affidabili e accurate le previsioni.

1.4.8 Data visualization

Se si guidi un auto o si voli su un aereo, le informazioni vitali su velocità, pressione dell'olio, temperatura e così via sono disponibili attraverso il cruscotto di fronte. Indicatori, luci rosse e verdi e contachilometri sono posizionati strategicamente in modo che con una rapida occhiata, senza perdere la concentrazione su dove si stia andando, si sappia se tutto stia procedendo correttamente, prendendo pertanto le decisioni di conseguenza. Proprio come i conducenti e i piloti fanno affidamento sui loro cruscotti per svolgere il loro lavoro, i manager si rivolgono sempre più ai cruscotti aziendali per aiutarli a gestire le loro organizzazioni. Le idee ei vantaggi sono praticamente gli stessi dell'esempio con il conducente: fornire ai manager un dashboard che su uno schermo ben progettato mostri le informazioni chiave di cui hanno bisogno per monitorare gli elementi di cui sono responsabili, quindi possono scoprire rapidamente i problemi e agire per aiutare a migliorare le prestazioni delle loro organizzazioni.

Il ruolo principale di un dashboard, per la cui descrizione è stato preso come stretto riferimento “Business Dashboards : A Visual Catalog for Design and Deployment” di Rasmussen e Nils H. [101], è pertanto il fornire ai manager un mezzo per monitorare, analizzare e talvolta annotare (ad esempio, spiegare le varianze in una scorecard incorporata), e ci sono diversi forti legami con la pianificazione e il budget:

- Visualizzazione, analisi e confronto storico con budget, previsioni e obiettivi.
- Monitoraggio e condivisione di strategie tra le unità aziendali.
- Monitoraggio delle cifre di allocazione delle risorse in base alle quali le unità aziendali possono proporre investimenti di fondi discrezionali in vari programmi e progetti.

Esistono molti modi in cui poter classificare i dashboard. Una soluzione può essere in tre tipologie principali:

- **Dashboard strategici:** obiettivi strategici che supportano l'allineamento organizzativo.

- **Dashboard tattici:** supportano la misurazione dei progressi in progetti o iniziative chiave.
- **Dashboard operativi:** supportano il monitoraggio di attività aziendali specifiche.

Il contenuto principale di una dashboard riguarda le metriche ed un loro sottinsieme: gli indicatori, conosciuti come KPI. In genere, questi ultimi sono rappresentati come un rapporto (percentuale) di una cifra effettiva rispetto ad una predefinita e vengono rappresentati graficamente in modo tale da consentire agli utenti di accorgersi immediatamente se si trovano sull'opportuno livello. In sintesi, i KPI ben progettati aiutano le organizzazioni a dedicare maggior tempo alle attività rilevanti, che guidano le prestazioni, e meno tempo alle non determinati. Poiché in genere i KPI sono le metriche più importanti di un'organizzazione, sono spesso il contenuto più prezioso delle dashboard stesse.

1.5 Software a confronto

Le piattaforme di Business Intelligence si sono evolute nella necessità di dare un senso ai dati aziendali, in gran parte numerici, riportandoli in tabelle strutturate di fogli di calcolo e database. Storicamente è stato poco agevole operare in tali contesti, a causa delle loro grandi dimensioni e dell'elevata dimensionalità. Tuttavia, l'invenzione della tabella pivot (introdotta per la prima volta in Lotus Improv nel 1991 e resa popolare da Microsoft Excel dal 1994) ha dato agli utenti la possibilità di esplorare i riepiloghi tabulari di queste tabelle ruotando interattivamente tra diverse aggregazioni numeriche (ad esempio, somma, media) per sottoinsiemi di righe e colonne. Nel 2000 Polaris ha esteso le tabelle pivot per consentire riepiloghi grafici di grandi database multidimensionali, ciò ha gettato le basi del prodotto software Tableau. Microsoft ha successivamente esteso Excel con funzionalità simili, pervenendo nel 2015 al rilascio di Microsoft Power BI come prodotto indipendente. L'utilizzo delle piattaforme BI è ora mainstream nel mondo degli affari e l'adozione è in crescita nelle sfere pubbliche della scienza, dell'ingegneria, dell'istruzione e di governo. Le piattaforme BI presentano i dati come "dashboard" di più

visualizzazioni collegate che riepilogano e consentono il filtraggio interattivo di un set di dati comune. Tuttavia, gran parte dei dati rilevanti per le organizzazioni moderne non sono sotto forma di tabelle numeriche strutturate, bensì sotto forma di testo e metadati non strutturati, distribuiti su documenti, social media e web. Anche se la portata di tali dati lo rende un candidato per l'analisi del dashboard, all'inizio del 2016 ancora nessuna delle principali piattaforme di BI supportava tali casi d'uso di dati non strutturati [103].

Un servizio di non poca rilevanza per il confronto delle prestazioni, e non solo, dei software attinenti alla business intelligence, è offerto dalla realtà Gartner, che realizza per ogni annata una comparazione tra i diversi prodotti presenti sul mercato, sulla base delle valutazioni e recensioni dei singoli clienti. La Gartner si attiene ad un particolare report denominato “Magic Quadrant”. Tale visualizzazione si dispone su due dimensioni: l’abilità di esecuzione e la completezza visiva. In relazione a tale spiegazione del mercato, a seconda delle rispettive grandezze, è possibile distinguere quattro tipologie di attori: niche players, visionaries, challengers e leaders [105].

Figure 1. Magic Quadrant for Analytics and Business Intelligence Platforms

Source: Gartner (February 2020)

Figura 9.

Di seguito svengono riportate le valutazioni di Gartner, sintetizzate da zerounoweb.it [104], riguardo le quattro soluzioni all'interno del quadrante "Leaders":

- **Power BI:** la soluzione di Microsoft, disponibile SaaS nel cloud di Azure o on-premise, gode della posizione migliore nel quadrante.
    - **Punti di forza:** la grande diffusione, dovuta sia ad una semplice ed istintiva accessibilità, sia al fatto che sia incluso in Office 365 E5. Le funzionalità del prodotto, grazie anche, tra l'altro, all'introduzione di tecnologie di

augmented analytics, ML automatizzato. La completezza della vision, in quanto Microsoft sta continuando a investire in nuove funzioni.

  - **Punti deboli:** la versione on premise presenta lacune funzionali sul fronte, per esempio, delle dashboard, dello streaming analytics, delle capacità NLP. Lavora solo su Azure, non offre la possibilità di scegliere un'altra soluzione IaaS. Difficoltà con SAP BX e HANA e generalmente preferiscono caricare i dati in Power BI.

- **Tableau:** nell'agosto 2019 Salesforce ha completato l'acquisizione di Tableau permettendone l'incremento delle potenzialità del vendor stesso; in particolare, Salesforce sostiene Tableau in 3 aree chiave per le soluzioni di Business Intelligence e Analytics ovvero AI, Cloud e embedded analytics.
  - **Punti di forza:** entusiasmo e soddisfazione degli utenti che dichiarano un'esperienza decisamente positiva e al di sopra della media utilizzando questa soluzione. Facilità di esplorazione visiva e manipolazione dei dati, attinti, tra l'altro, da un'ampia gamma di fonti. Possono essere organizzati in modo semplice, per esempio creando gruppi, contenitori e gerarchie.
  - **Punti deboli:** i rischi riconducibili a un mercato in rapida evoluzione, tutti i competitors si stanno muovendosi velocemente verso l'era dell'analisi aumentata ma comunque, sinora, Tableau - secondo Gartner - ha compiute mosse corrette. Funzionalità di governance deboli, nonostante siano stati miglioramenti nella versione 2019. Gli utenti hanno dichiarato di apprezzare l'opzione a basso costo Tableau Viewer e stanno guardando con attenzione ai prezzi degli abbonamenti nei quali non sono contemplate però alcune componenti aggiuntive quali Server management e Data management.

- **Quilk:** tra le soluzioni di Business Intelligence e Analytics di Qlik la principale è Qlik Sense eseguita sull'esclusivo Qlik Associative Engine, il motore cognitivo cui sono state aggiunte funzionalità AI / ML. L'acquisizione di Attunity, il cui prodotto

rimane autonomo, amplia le capacità di integrazione dei dati dell'ecosistema di questo vendor.

  - **Punti di forza:** flessibilità di implementazione on premise e con qualsiasi cloud provider, garantita anche la possibilità di approcci combinati, è da sempre uno dei plus di Qlik. L'ampiezza delle funzionalità della piattaforma, per abbracciare varie fasi del ciclo di vita dell'analytics; in particolare, sebbene Qlik Data Integration Platform (precedentemente Attunity) sia un'offerta autonoma, mette a disposizione potenti funzionalità di integrazione e spostamento dei dati sotto l'ombrello di Qlik. L'alfabetizzazione dei dati: uno degli obiettivi del vendor è proprio quello di creare e arricchire una cultura data driven che sfrutti la sua tecnologia che permette capacità associativa, oltre a funzionalità aumentate , al fine di scoprire automaticamente informazioni che potrebbero sfuggire.
  - **Punti deboli:** momento storico; oggi Qlik è di nuovo in cerca di nuove persone dopo un 2019 di acquisizioni tecnologiche e rinnovi di prodotto ma non si rileva particolare slancio sul mercato. Nonostante l'impegno di Qlik nella fornitura di supporto e risorse dedicate ai clienti che passano da QlikView a Qlik Sense tale passaggio, negli utenti, desta preoccupazioni. Anche se Qlik Sense è stato progettato per supportare il self-service per la visualizzazione dei dati, gli utenti interpellati da Gartner hanno utilizzato dashboard parametrizzate; l'esperienza associativa possibile con la tecnologia Qlik offre metodi alternativi per scoprire gli insight che riducono le necessità di dover ricorrere al self service.

- **Thoughtspot:** il suo innovativo approccio alla ricerca continua a distinguersi sul mercato, in particolare per la possibilità di porre domande analiticamente.
  - **Punti di forza:** utilizza il NLP come principale mezzo di interrogazione, analisi su grandi quantità di dati (un terzo dei suoi utenti contribuisce a oltre un terabyte di dati). SpotIQ, grazie alla sua capacità di augmented analytics, consente di scoprire anomalie, correlazioni e analisi comparative

senza necessità di coding. Sta inoltre dando la priorità alla costruzione nella sua piattaforma di esperienze di consumo simili a Facebook. Le priorità includono feed continui di contenuti correlati, rilevamento automatico delle anomalie e avvisi proattivi. Nonostante le dimensioni ridotte di ThoughtSpot, il mercato riconosce il valore della sua tecnologia soprattutto sul fronte NLP e analisi aumentata.

- **Punti deboli:** il software ThoughtSpot tipicamente integra altri prodotti e non copre l'intero spettro dei requisiti delle soluzioni di Business Intelligence e Analytics. Nello specifico, per esempio, i dati devono essere preparati e puliti utilizzando strumenti di terze parti. In pratica, pochi utenti vedono ThoughtSpot come unico interlocutore per queste piattaforme. Il vendor è in crescita, il suo ecosistema di partner sta arricchendosi grazie a nuovi investimenti compiuti nel 2019. La corretta implementazione del software ThoughtSpot richiede la preparazione dei dati e la mappatura dei dati in anticipo. Questo in genere richiede competenze e impegno IT. Inoltre, il prodotto offre contenuti precostruiti limitati per domini verticali e specifici. I clienti devono quindi creare le proprie applicazioni per particolari aree funzionali.

CAPITOLO II

Power BI

2.1 Descrizione

Il software Power BI è un prodotto fornito dallo sviluppatore Microsoft, ideato da Thierry D'Hers e Amir Netz e realizzato da Ron George nel 2010, inizialmente con la denominazione di Project Crescent. Fu rinominato con l'odierno nome solo successivamente e, inizialmente, era finalizzato alla sola fornitura di tre componenti aggiuntivi per l'applicazione Microsoft Excel: Power Query, Power Pivot e Power View. Nel 2015 è stato successivamente rilasciato con implementazione di ulteriori caratteristiche, quali la connettività dei dati e le opzioni di sicurezza, quest'ultime mediante Power BI Gateway [106, 113].

Power BI è un servizio di analisi business, una raccolta pertanto di funzioni, app e connettori che interagiscono per trasformare le origini dei dati, non correlate, in un insieme di informazioni coerenti, mediante quindi una vista unificata dei dati più importanti, facilmente comprensibile e interattiva. I dati possono provenire da diverse fonti come un foglio di calcolo di Excel o una raccolta di data warehouse basati sul cloud o ibridi locali e la loro connessione e integrazione sono semplificate. Questi vengono visualizzati e analizzati in modo da permette il monitoraggio dell'integrità del business in tempo reale, anche con diverse angolature, mediante un dashboard ed i relativi report. Power BI consente quindi la produzione di informazioni critiche per l'intera organizzazione, riuscendo a riunire le metriche più rilevanti. In ultimo stadio è possibile la condivisione con gli utenti, altro personale o semplicemente con che ne ha necessità [106, 113].

Figura 10.

2.2 Struttura e funzioni

I diversi servizi offerti dal software Power BI, differiscono per finalità e complessità ma contribuiscono in egual misura al realizzo di un progetto di business intelligence, riconducibile a diversi step: si inizia con i diversi passi comprendenti l'estrapolazione, la modifica e la compattazione delle informazioni, per terminare con la condivisione di queste, nelle più e varie piattaforme disponibili. I servizi offerti pertanto, non sono fruibili da un'unica sezione bensì, da precise e distaccate parti. Il prodotto della Microsoft infatti, è stato strutturato e suddiviso in tre distaccati ambiti, ognuno dei quali dedito a precise funzioni e contesti. Si distinguono le seguenti componenti: Power BI Desktop, Power BI Service e Power BI Mobile, ognuna con le loro individuali caratteristiche e funzionalità [106, 107].

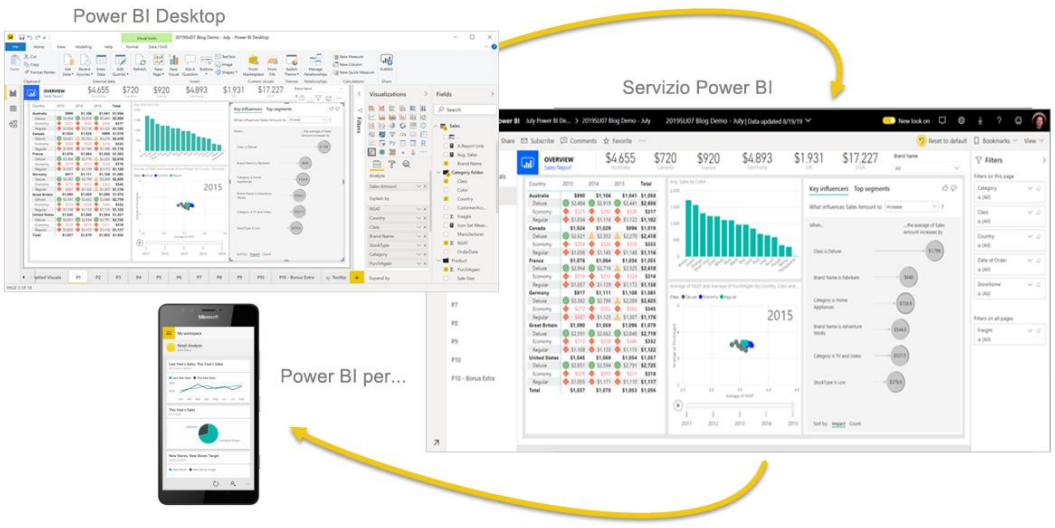


Figura 11.

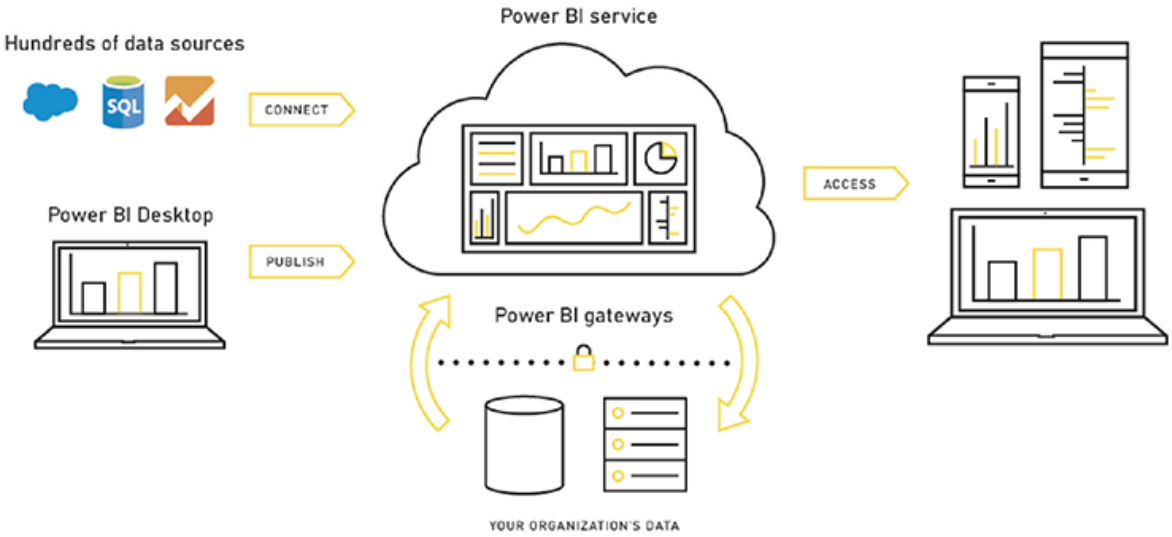


Figura 12.

2.2.1 Power BI Desktop

Power BI Desktop, da quanto estrapolato per tutto il corrente e seguenti capitoli dal "Microsoft Power BI Quick Start Guide : Build Dashboards and Visualizations to Make Your Data Come to Life" di Knight e Devin [114], è un'applicazione gratuita e una delle parti complementari del software Power BI, che viene installata nel computer locale e consente di connettersi ai dati, trasformandoli e visualizzandoli. Power BI Desktop permette di collegarsi a molte e diverse origini dati, di combinarle (*modellazione*) integrandole in un modello di dati che consente successivamente di creare oggetti visivi e raccolte di oggetti come i report, condivisibili con altri utenti all'interno dell'organizzazione (quest'ultimo anche mediante Power BI Service con il quale poter inoltre realizzare la dashboard). La maggior parte dei progetti di business intelligence infatti, si basano su Power BI Desktop per la creazione di report e sul *servizio Power BI* per la condivisione di report con altri utenti.

Caratteristica peculiare di Power BI Desktop è che le diverse funzioni a disposizione, sono disponibili in due ulteriori distinte sezioni, ognuna della quali riguardante un preciso ambito e quindi step di un progetto di business intelligence.

- **Data preparation**: in questa sezione è possibile la connessione ai dati, la trasformazione e la pulizia di questi per la creazione di un modello mediante relazioni. È possibile accedervi, una volta eseguito il software Power BI Desktop, dalla prima finestra di accesso, per mezzo del comando “Trasforma dati”.

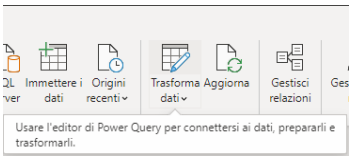


Figura 13.

- **Data model:** suddiviso in tre viste, una di queste (che prende il nome di data view) permette la creazione di oggetti visivi, come grafi, diagrammi e ulteriori visualizzazioni, che forniscono le rappresentazioni grafiche dei dati. In questi è possibile selezionare i campi desiderati, aggiungere filtri, formattare con colori personalizzati, gradienti e altre opzioni; è quindi possibile la creazione di report, che raccolgono oggetti visivi in una o più pagine. In ultimo è fruibile la funzione di condivisione delle viste realizzate. La seconda vista riguarda la data modelling dove è possibile visualizzare le tabelle, le misure e altri dati usati nel modello di dati associato al report, nonché trasformare i dati per un uso ottimale nel modello del report stesso. Terza e ultima vista è la data relationship che dispone degli strumenti per le relazioni tra le differenti tabelle nel modello di dati.

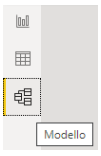


Figura 14.

Gli utenti che si occupano di queste attività spesso sono chiamati *analisti di dati* (talvolta solo *analisti*) o professionisti di Business Intelligence (spesso chiamati anche *creatori di report*). Tuttavia, senza associarsi a tali figure, si può adoperare Power BI Desktop per la semplice creazione di report efficaci o per eseguire il pull dei dati da origini diverse e creare modelli di dati condivisibili con i colleghi e le organizzazioni.

2.2.1.1 Connessione ai dati

Power BI è meglio conosciuto per le chiare e colarate visualizzazioni dei dati, oltre che per le funzionalità del dashboard di cui dispone. Tuttavia, prima di poter iniziare a creare rapporti, occorre connettersi alle origini dati necessarie. Questa operazione è possibile nella sezione della data preparation all'interno della quale uno sviluppatore ha più di 80 connettori dati univoci tra cui scegliere, che vanno dai tipi di file tradizionali ai motori di database, soluzioni per big data, origini cloud, dati archiviati su una pagina Web e altri provider SaaS.

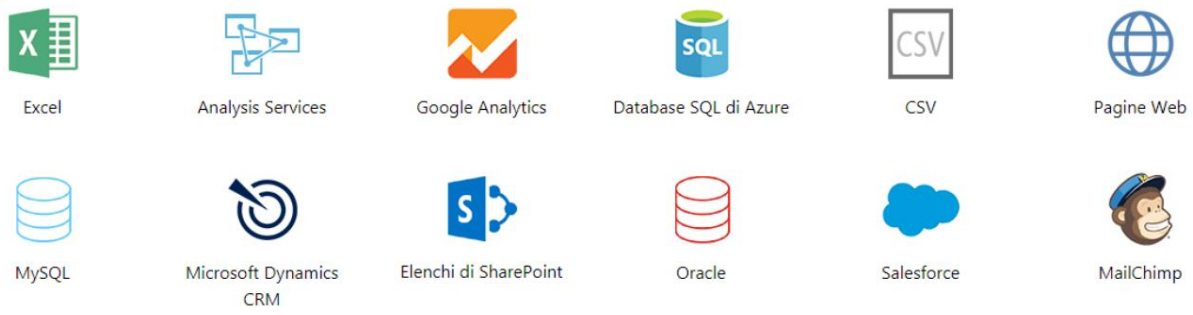


Figura 15.

Quando si stabilisce una connessione a un'origine dati, è possibile che venga presentata una delle tre diverse opzioni su come questi devono essere trattati: importa, DirectQuery o Live Connection.

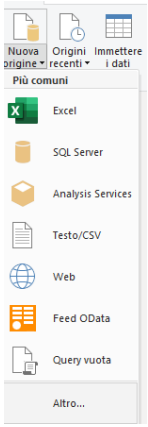


Figura 16.

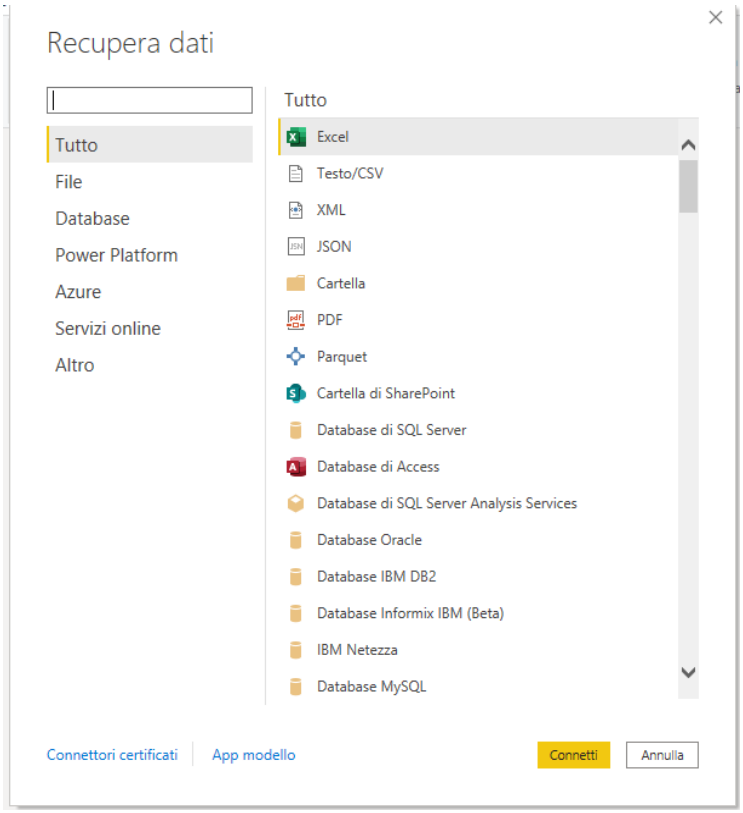


Figura 17.

- **L'importazione dei dati**: è l'opzione più comune e consiste nell'estrazione fisica delle righe di dati dall'origine selezionata, le quali verranno archiviate in memoria all'interno di Power BI. Power BI Desktop utilizza un metodo speciale per l'archiviazione dei dati, noto come xVelocity, che è una tecnologia in memoria che non solo aumenta le prestazioni dei risultati delle query, ma può anche comprimere notevolmente la quantità di spazio occupata dalla soluzione Power BI. In alcuni casi, la compressione che si verifica può anche ridurre lo spazio richiesto su disco fino a un decimo della dimensione dell'origine dati originale. Il motore xVelocity utilizza un'istanza invisibile locale di SQL Server Analysis Services (SSAS) per fornire queste funzionalità in memoria. L'applicazione del Data Gateway è inoltre un requisito per pianificare gli aggiornamenti dei dati per le origini che utilizzavano l'opzione di importazione.
- **DirectQuery:** Power BI, mediante questa opzione, consente di connettersi direttamente a un'origine dati in modo che nessun dato venga importato o

copiato in Power BI Desktop. Poiché nessun dato viene importato in Power BI Desktop, ciò significa che è meno importante quanto sia potente il proprio laptop personale perché tutti i risultati delle query vengono, in questa modalità, elaborati sul server di origine anziché sul proprio laptop appunto. Significa anche che non è necessario aggiornare i risultati in Power BI perché tutti i report progettati puntano sempre a una versione live dell'origine dati. L'abilitazione di questa funzionalità può essere eseguita semplicemente selezionando DirectQuery durante la configurazione di un'origine dati.

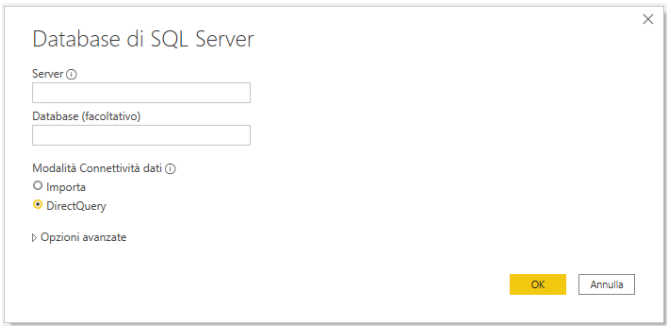


Figura 18.

Anche in questo contesto è necessario il Data Gateway, in quanto permette di inviare i dati in locale nel cloud. Nell'utilizzo dell'opzione DirectQuery ci sono però delle limitazioni. La prima tra tutte è che non tutte le origini dati supportano la DirectQuery stessa. Alcune di queste sono: Amazon Redshift, Azure HDInsight Spark, Azure SQL Database, Azure SQL Data Warehouse, Google BigQuery, IBM Netezza, Impala (versione 2.x), Database Oracle (versione 12 e successive), SAP Business Warehouse Application Server, SAP Business Warehouse Message Server, SAP HANA, Snowflake, Spark (Version 0.9 and above), SQL Server, Teradata Database, Vertica. A seconda dell'origine dati scelta è possibile inoltre che le prestazioni delle query siano più lente quando si utilizza DirectQuery rispetto all'opzione di importazione dati predefinita. Occorre tenere presente infatti che quando l'opzione di importazione è selezionata, sfrutta un motore di archiviazione in memoria altamente sofisticato. Quando si seleziona DirectQuery, le prestazioni dipendono invece dal tipo di origine scelto. Un'altra limitazione degna di nota è che non tutte le funzionalità di Power BI, come nell'editor, sono

supportate quando si sceglie l'opzione DirectQuery. Ad esempio alcune funzioni DAX non sono disponibili.

- **Live Connection:** è molto simile all'opzione DirectQuery. Infatti, quando si usa una connessione dinamica, nessun dato viene effettivamente importato in Power BI. La soluzione invece punta direttamente all'origine dati sottostante e sfrutta Power BI Desktop semplicemente come uno strumento di visualizzazione dei dati. La differenza tra DirectQuery e Live Connection, è che con quest'ultimo, in genere, non si hanno problemi di prestazioni perché è supportato solo dai seguenti tipi di origini dati: SQL Server Analysis Services Tabular, SQL Server Analysis Services Multidimensional, Service Power BI. Il motivo per cui le prestazioni non soffrono con queste origini dati è perché queste ultime usano lo stesso motore xVelocity di Power BI o un altro motore di archiviazione ad alte prestazioni. Per impostare la propria connessione in tempo reale a una di queste origini, è possibile scegliere il database di SQL Server Analysis Services dall'elenco delle origini.

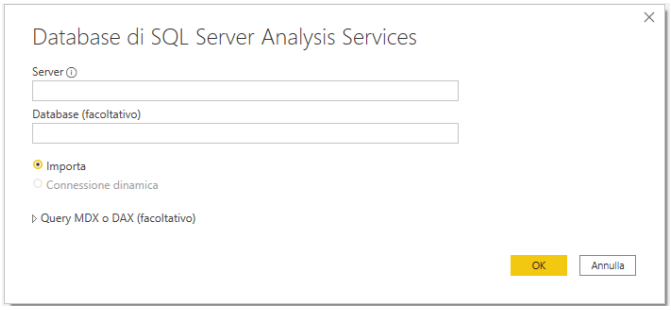


Figura 19.

2.2.1.2 Trasformazioni

All'interno di qualsiasi progetto di BI, è essenziale che i dati con cui si lavora siano stati adeguatamente puliti, per ottenere risultati accurati sui report e sulle dashboard. L'applicazione di regole aziendali di pulizia dei dati, note anche come trasformazioni, è il metodo per correggere dati imprecisi o non corretti, ma tale processo può spesso essere la parte più dispendiosa, in termini di tempo, di qualsiasi soluzione di BI aziendale. Tuttavia, le funzionalità di trasformazione dei dati integrate in Power BI, sono molto

potenti e intuitive. Utilizzando l'Editor di Power Query, le attività che in genere sarebbero difficili o che richiedono molto tempo in uno strumento di BI aziendale sono semplici, come fare clic con il pulsante destro del mouse su una colonna e selezionare la trasformazione appropriata per il campo. Durante l'interazione con l'interfaccia utente in questo editor, dietro le quinte viene scritto automaticamente un linguaggio chiamato 'M'.

L'editor di Power Query è lo strumento principale per applicare le trasformazioni dei dati e i processi di pulizia alla propria soluzione. Quando viene aperto l'editor di Power Query, si nota che ha un proprio ambiente separato in cui lavorare.

L'applicazione delle trasformazioni dei dati nell'editor di Power Query può essere una cosa sorprendentemente semplice da fare. Tuttavia, ci sono poche cose da considerare quando si inizia questo processo; il primo è che esistono diversi modi per risolvere un problema e che, come detto, ad ogni clic effettuato nell'editor di Power Query avviene l'automatica conversione in un linguaggio ad hoc chiamato 'M'. La maggior parte delle trasformazioni pertanto, possono essere realizzate semplicemente interagendo con l'interfaccia utente dell'editor di Power Query ma, per problemi aziendali più complessi, ci sono buone probabilità che si ricorra alla modifica delle query M, scritte direttamente dall'editore. Infine, ultima considerazione importante da fare è che tutte le trasformazioni create nell'editor vengono archiviate in uno specifico riquadro denominato "passaggi applicati". La sezione passaggi applicati ha molte funzionalità:

- **Eliminazione delle trasformazioni**: se si commette un errore e si deve annullare un passaggio, è possibile fare clic sul pulsante "elimina" accanto ad un passaggio.
- **Modifica delle trasformazioni**: questa operazione può essere eseguita con qualsiasi passaggio che abbia accanto un'icona a forma di ingranaggio.
- **Modifica dell'ordine delle trasformazioni:** se ci si rende conto che è meglio eseguire un passaggio prima di un altro, si può modificare l'ordine di esecuzione dei passaggi.

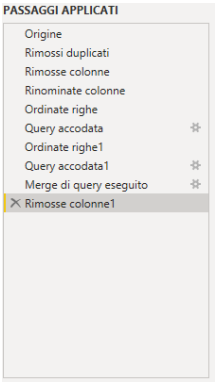


Figura 20.

In questo contesto, si hanno a disposizione una notevole quantità di trasformazioni realizzabili sui dati delle diverse fonti con le quali si sia stabilita una connessione. Spesso, per ogni funzionalità offerta da Power BI, si hanno diversi modi per accedervi; nella maggior parte dei casi però, queste sono accessibili direttamente dalla barra di lavoro al di sopra del set di dati:

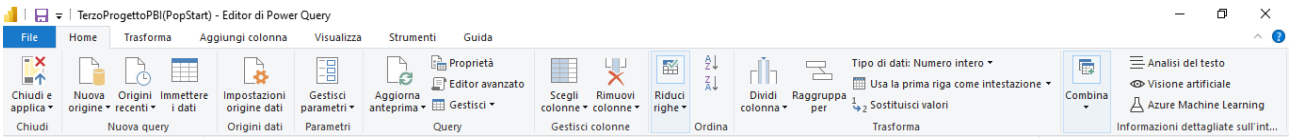


Figura 21.

- **Usare la prima riga come intestazioni:** l'organizzazione dei nomi o delle intestazioni delle colonne è spesso una delle prime attività importanti durante l'organizzazione del set di dati. Fornire dei nomi delle colonna pertinenti, rende molti dei processi a valle, come la creazione di report, molto più semplici. Spesso, le intestazioni di colonna vengono importate automaticamente dall'origine dati ma, a volte, si può lavorare con un'origine dati più complessa, che rende difficile per Power BI acquisire le informazioni sull'intestazione di colonna.

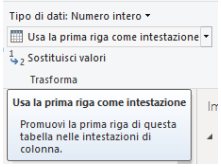


Figura 22.

- **Rimozione colonne:** spesso, le origini dati alle quali ci si collega, includono molte colonne che non sono necessarie per la soluzione del proprio progetto. È importante rimuovere queste colonne non necessarie dal set di dati poiché, in quanto inutilizzate, occupano inutilmente spazio all'interno del modello di dati. Come accennato, anche in questo caso esistono diversi metodi per rimuovere le colonne nell'editor di Power Query.

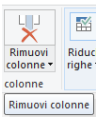


Figura 23.

- **Cambiamento del formato:** la definizione corretta dei tipi di dati delle colonne nelle prime fasi del processo di pulizia dei dati può aiutare a determinare il tipo di valori con cui si sta lavorando. L'editor di Power Query dispone di vari tipi di dati numerici, di testo e di data e ora tra cui scegliere. In ogni caso l'editor di Power Query interpreta automaticamente i formati delle colonne.

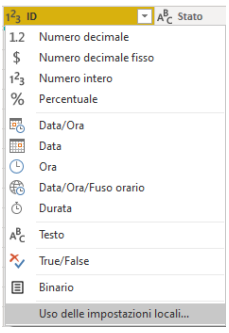


Figura 24.

- **Aggiunta colonne:** un'opzione che può rendere complesse delle semplici trasformazioni di dati è la funzione chiamata “Colonna da esempi” dalla sezione “Aggiungi Colonna”. Con questa infatti si fornisce all'editor di Power Query un esempio di come si vorrebbe che fossero i propri dati e quindi determinare automaticamente quali trasformazioni sono necessarie per raggiungere il proprio obiettivo.

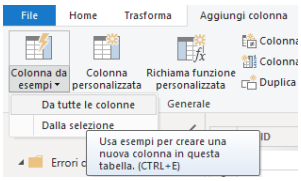


Figura 25.

- **Colonne condizionali**: l'utilizzo di questa funzionalità in Power Query è un ottimo modo per aggiungere nuove colonne alla query che seguono le istruzioni logiche if / then / else. Questo concetto di if / then / else è comune a molti linguaggi di programmazione, comprese le formule di Excel.

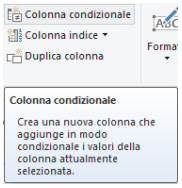


Figura 26.

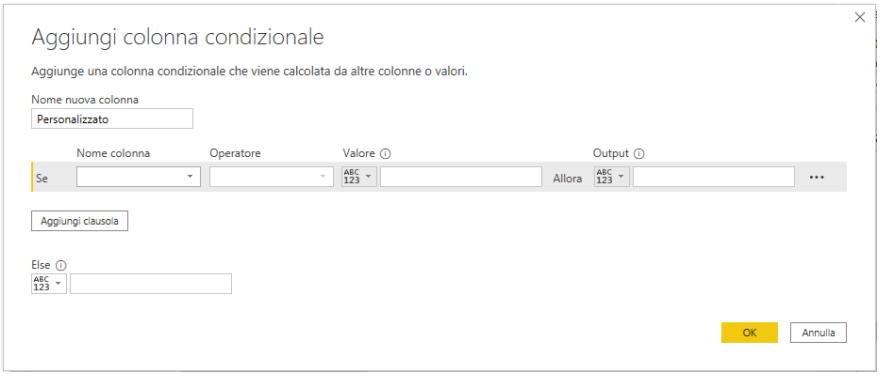


Figura 27.

- **Fill Down:** è una trasformazione piuttosto unica nel modo in cui opera. Selezionando Riempi in basso su una particolare colonna, un valore sostituirà tutti i valori nulli sottostanti fino a quando non appare un altro non nullo. Si può anche usare Fill Up per realizzare il contrario ovvero sostituire tutti i valori nulli con un valore non nullo.

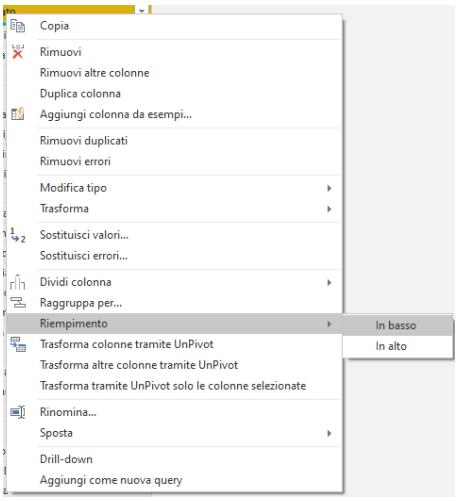


Figura 28.

- **Unpivot:** questa trasformazione è molto potente e consente di riorganizzare il set di dati in un formato più strutturato per la Business Intelligence. Permette di trasformare le intestazioni delle colonne selezionate in valori di riga e i dati in quelle stesse colonne in una riga corrispondente. Con questa selezione, tutte le nuove colonne che possono essere aggiunte all'origine dati verranno automaticamente incluse nella trasformazione unpivot.

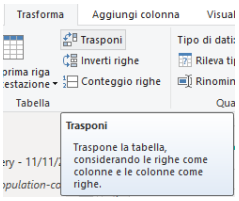


Figura 29.

- **Merge:** un requisito comune quando si creano soluzioni BI è la necessità di unire due tabelle insieme per formare un nuovo risultato che includa alcune colonne di entrambe le tabelle in una singola query. Power BI rende questa attività molto semplice con la funzionalità “Unisci query”. L'utilizzo di questa funzione richiede la selezione di due tabelle e quindi la determinazione della colonna o delle colonne che costituiranno la base per l'unione delle due query. Dopo aver determinato le colonne appropriate per il join, occorre selezionare il tipo di join:
    - **Left outer:** tutte le righe della prima tabella, matching delle righe della seconda tabella.

- **Right outer:** tutte le righe della seconda tabella, matching delle righe della prima tabella.
- **Full outer:** solo le righe in comune tra le due tabelle.
- **Inner:** tutte le righe di entrambe le tabelle.
- **Left anti:** solo le righe della prima tabella prive di corrispondenza con la seconda tabella.
- **Right anti:** solo le righe della seconda tabella prive di corrispondenza con la prima tabella.

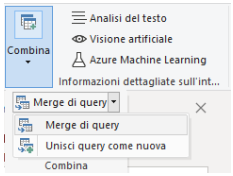


Figura 30.

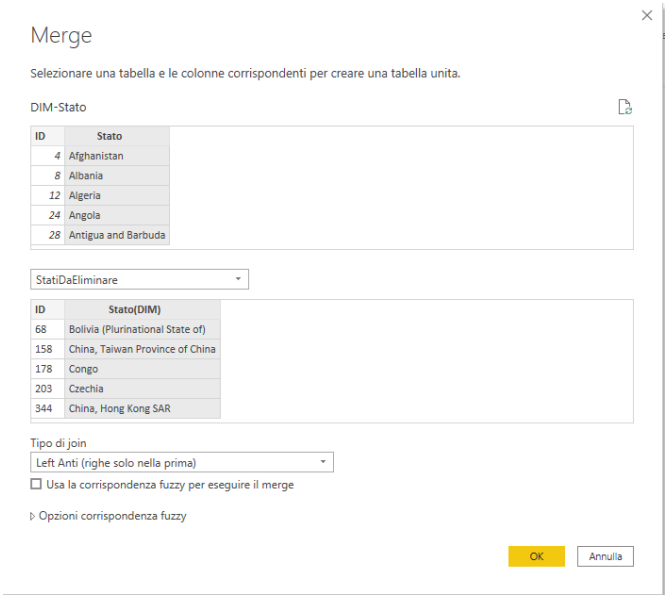


Figura 31.

- **Appending**: occasionalmente, è possibile lavorare con più set di dati che devono essere aggiunti l'uno all'altro. Uno scenario può essere quello di lavorare per un reparto di assistenza clienti, presso un'azienda che fornisce credito ai clienti. Vengono forniti regolarmente file .csv e .xlsx che forniscono riepiloghi dei reclami dei clienti relativi a carte di credito e prestiti. Si vuole poter analizzare entrambi

questi estratti di dati contemporaneamente ma, purtroppo, i reclami relativi alla carta di credito e al prestito sono forniti in due file separati. In questo caso è conveniente adoperare un'operazione di accodamento su questi due stessi diversi file.

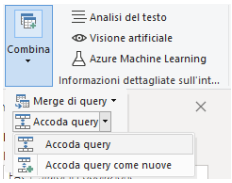


Figura 32.

2.2.1.3 Linguaggio R

R è un linguaggio di scripting molto rilevante che viene utilizzato principalmente per strumenti di analisi avanzati ma ha anche diversi punti di integrazione all'interno di Power BI. Una di queste integrazioni è la capacità di estendere i limiti dell'editor di Power Query e, pertanto, chiamare funzioni e librerie da R per adoperare funzioni che non sarebbero regolarmente possibili. Per usare il linguaggio R in Power BI, occorre prima installare una distribuzione R per eseguire i diversi script. Uno dei più utilizzati è Microsoft R Open, un progetto open source e gratuito.

2.2.1.4 Linguaggio M

L'editor di Power Query è l'interfaccia utente che permette di progettare e creare dalle importazioni di dati. Tuttavia, per ogni trasformazione che si applica all'interno di questo editor in realtà, corrisponde una query M scritta automatica da Power BI. Per soluzioni semplici, è improbabile che si abbia il bisogno di guardare la query M scritta ma ci sono casi, più complessi, in cui è necessario.

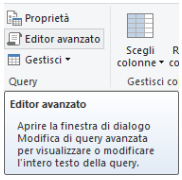


Figura 33.

Figura 34.

Alcuni dei principali aspetti riguardanti la sintassi e le caratteristiche del linguaggio M possono essere raccolte in questi concetti:

- **Espressione “Let”:** incapsula un insieme di valori o espressioni con nome da calcolare.
- **Espressione “In”:** stranamente, l'espressione “In” è in realtà un riferimento a ciò che verrà restituito dalla query nella sezione “Passaggi applicati” e pertanto nell'anteprima dell'editor di Power Query.
- **Espressioni o variabili denominate**: il nome assegnato a un insieme di operazioni in un’operazione. Questi nomi possono essere qualsiasi cosa, ma è necessario notare che se si desidera avere uno spazio nel nome di un passaggio, è necessario racchiuderlo tra # "". Ad esempio, se volessi che qualcosa si chiamasse “Operazione 1”, allora dovrei chiamare un'espressione # "Passaggio 1".
- **Funzioni M:** corrisponderebbero alle operazioni utilizzate per manipolare l'origine dati. Il linguaggio “M Query” generalmente esegue le sue funzioni come operazioni seriali, il che significa che ogni operazione viene eseguita una dopo l'altra in sequenza.

2.2.1.5 Data model

Storicamente, i progetti di Business intelligence focalizzati sulla creazione di modelli di dati potevano richiedere mesi e persino anni per svilupparsi, soprattutto quando si lavorava all'interno della rigida struttura e dei vincoli di un ambiente di business intelligence aziendale. Alcuni studi mostrano che circa il cinquanta per cento di tutti i progetti di BI falliscono e che questi progetti o non vengono completati o non mantengono i risultati promessi al loro completamento. Power BI Desktop, invece, è uno di quei software che ha offerto un approccio molto più agile alla creazione del modello di dati e, al posto di mesi o anni, la creazione di un modello di dati può richiedere anche ore o al massimo giorni.

Si potrebbe sostenere che la creazione di relazioni sia di fatto l'elemento più importante di Power BI Desktop. È questo processo, ovvero la costruzione di relazioni, d'altronde, che fa funzionare il resto delle operazioni previste; il filtraggio automatico di oggetti visivi e report, la facilità con cui è possibile creare misure DAX e la capacità di connettere rapidamente origini dati disparate, sono tutti possibili grazie alle relazioni correttamente create nel modello di dati. Alle volte, Power BI Desktop crea automaticamente le relazioni e pertanto risulta importante verificarle per garantirne l'accuratezza.

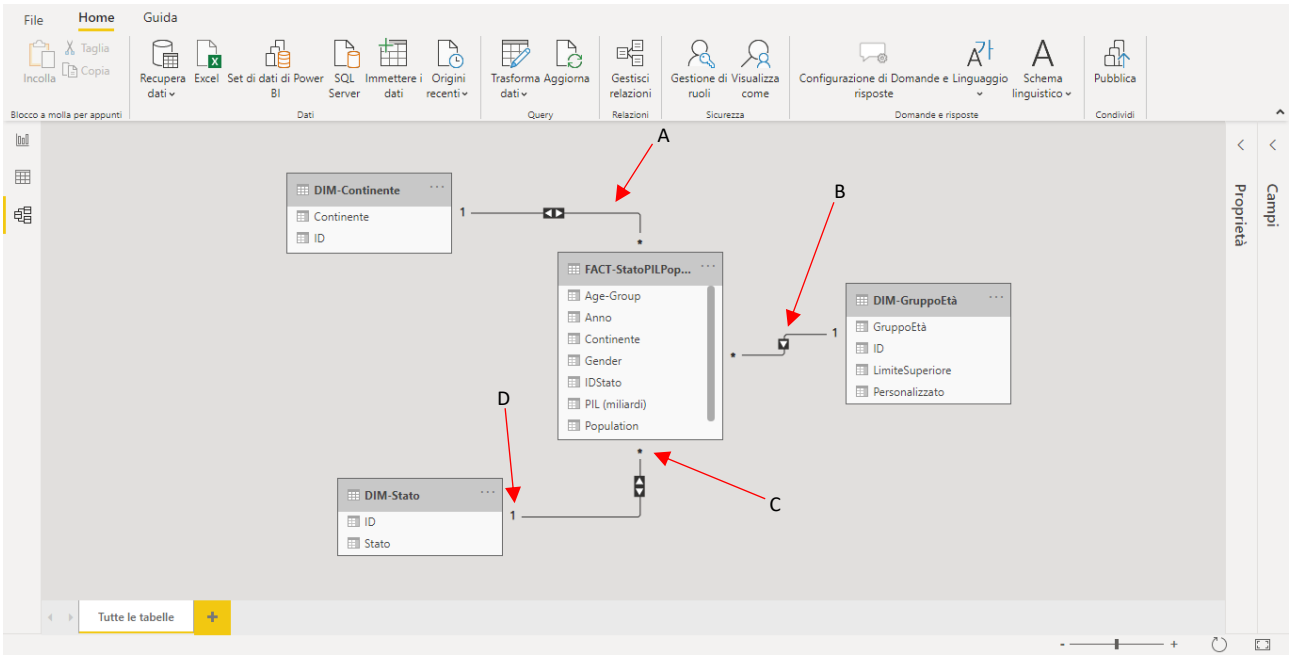


Figura 35.

Nella figura precedente, riscontrabile alla finestra d'apertura di Power BI Desktop nell'apposita sezione "relazioni", si nota la composizione di una stessa relazione tra diversi dati, rappresentata da quattro principali aspetti:

- **A = Relazione**: la linea tra due tabelle rappresenta l'esistenza di una relazione.
- **B = Direzione**: la freccia indica la direzione in cui avverrà il filtraggio
- **C = Cardinalità "molti":** indicato dal simbolo *, delinea che la tabella FactStatoPilPopolazione sia, nella relazione, la tabella nella quale più sue righe abbiano una medesima corrispondenza dall'altro lato della connessione.
- **D = Cardinalità "uno":** indicato dal numero 1, delinea che ogni riga della tabella DIM-Stato può avere più corrispondenze dall'altro lato della relazione.

  La prima cosa da fare dopo aver importato i dati è verificare che tutte le relazioni rilevate automaticamente siano state create correttamente. Dalla barra multifunzione di modellazione, si seleziona Gestisci relazioni:

Al fine pertanto di modificare o correggere le diverse relazioni, si hanno a disposizione diversi modi. Uno di questi è per mezzo dell'apposito comando detto appunto "Gestisci relazioni", che consente di avere una visione in aggregato delle diverse connessioni tra le tabelle realizzate nell'editor.

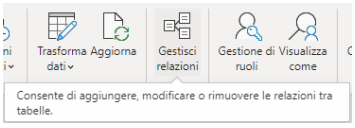


Figura 36.

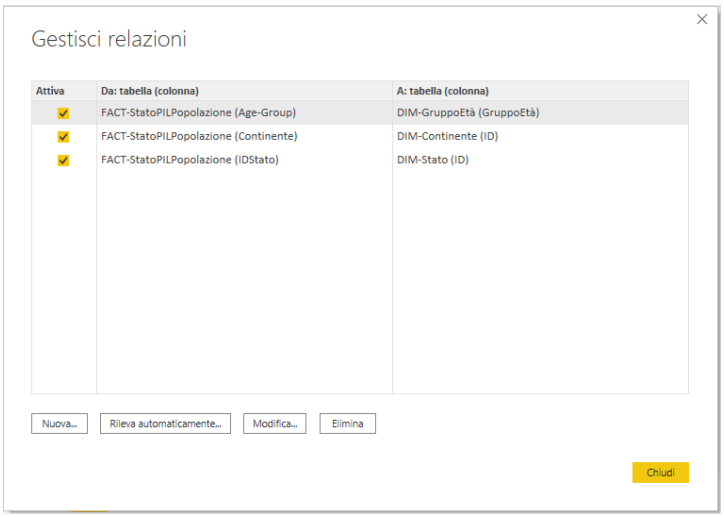


Figura 37.

Da qui si giunge all'apposito editor con il quale è possibile gestire o eseguire le seguenti operazioni:

- Visione delle relazioni correnti nel modello di dati
- Creazione di una nuova relazione

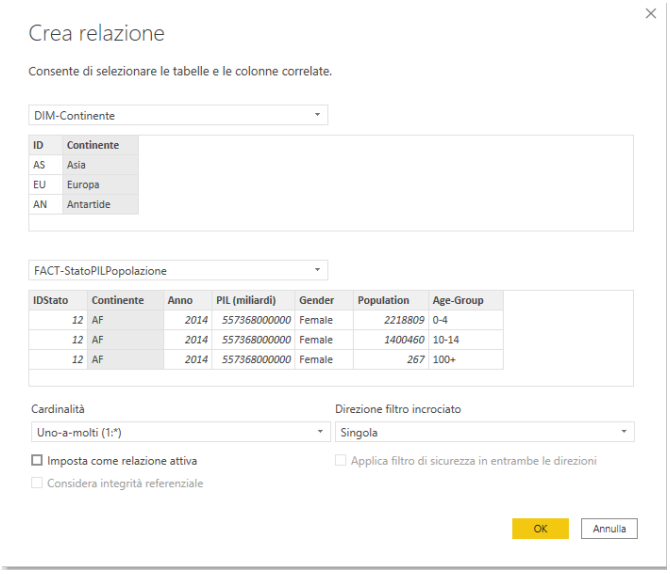


Figura 38.

- Modifica delle relazioni esistenti

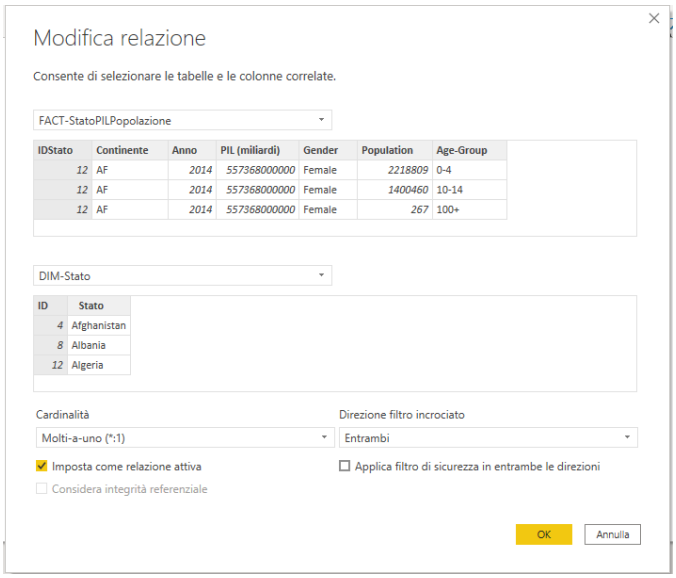


Figura 39.

- Eliminazione di una relazione

2.2.1.6 Miglioramento delle prestazioni

I miglioramenti dell'uso sono quegli accorgimento che possono migliorare significativamente l'esperienza complessiva dell'utente durante l'interazione con il modello di dati. Alcuni di questi sono:

- **Nascondere tabelle e colonne:** alcune tabelle sono disponibili nel modello dati semplicemente come supporto e non verrebbero utilizzate in un report. Ad esempio, si potrebbe avere una tabella per supportare le relazioni molti-a-molti, l'allocazione ponderata o anche la sicurezza dinamica. Allo stesso modo, alcune colonne sono necessarie per creare relazioni nel modello di dati ma non aggiungerebbero alcun valore quando vengono aggiunte a un report. Le tabelle o le colonne che non verranno utilizzate per scopi di reportistica, dovrebbero essere nascoste dalla visualizzazione report per ridurre la complessità e migliorare le prestazioni.

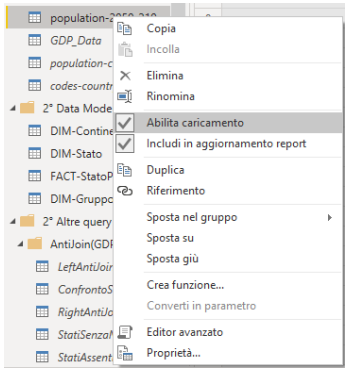


Figura 40.

- **Ridenominazione di tabelle e colonne:** è un passaggio importante per semplificare l'utilizzo del modello di dati. I diversi reparti spesso hanno termini diversi per la stessa entità, quindi è importante considerare più reparti quando si rinominano gli oggetti. È pertanto rilevante tenere a mente il destinatario dei rapporti, ovvero gli utenti finali e i consumatori, quando si rinominano tabelle e colonne.

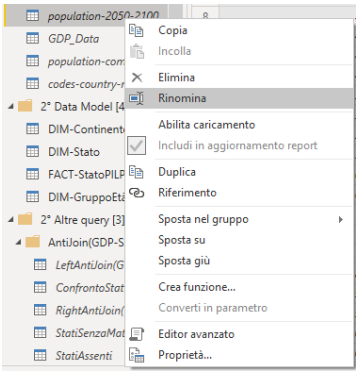


Figura 41.

- **Visualizzazione di una colonna in base a un'altra:** spesso si desidera visualizzare il nome di una colonna ma ordinarla in base ad un'altra. Ciò è possibile mediante l'apposito comando.

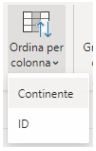


Figura 42.

- **Impostazione della categoria di dati:** Power BI fa alcuni presupposti sulle colonne in base ai tipi di dati, ai nomi e alle relazioni nel modello di dati. Questi presupposti vengono utilizzati nella visualizzazione dei report durante la creazione delle visualizzazioni, per migliorare l' esperienza utente. Quando Power BI rileva una colonna con valori numerici, infatti, viene assegnata un'aggregazione predefinita che può però essere in corso modificata. La classificazione dei dati consente come detto di migliorare l'esperienza dell'utente e di perfezionare la precisione. Sono disponibili diverse opzioni per la categorizzazione dei dati ma l'utilizzo più comune è per la classificazione dei dati geografici.

Figura 43.

- Creazione delle gerarchie: può fornire diversi vantaggi chiave. Alcuni di questi vantaggi sono elencati di seguito:
    - Le gerarchie organizzano gli attributi e mostrano le relazioni nei dati
    - Le gerarchie consentono una facile interattività
    - Le gerarchie aggiungono un valore analitico significativo al livello di visualizzazione attraverso il drill down e il rollup dei dati.
    - Le gerarchie memorizzano informazioni sulle relazioni nei dati, che gli utenti potrebbero non conoscere altrimenti.

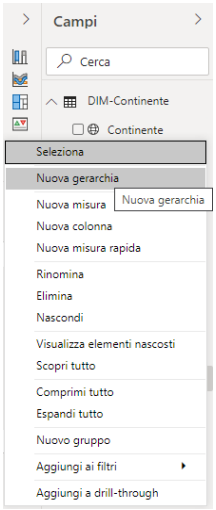


Figura 44.

2.2.1.7 Linguaggio DAX

Data Analysis Expressions (DAX) è un linguaggio di espressioni che ha fatto il suo debutto nel 2010 con il rilascio di Power Pivot in Excel. Gran parte di DAX è simile alle funzioni di Excel e quindi l'apprendimento di DAX è una transizione facile per gli utenti di Excel e gli utenti esperti. Le formule DAX includono funzioni, operatori e valori per eseguire query e calcoli avanzati sui dati presenti in tabelle e colonne correlate nei modelli di dati tabulari. Power BI Desktop dispone inoltre di IntelliSense, una funzionalità incorporata che contribuisce al completamento automatico del codice mentre e che, pertanto, consente di esplorare e scoprire le diverse funzioni del linguaggio DAX. Per sfruttare IntelliSense, è sufficiente iniziare a digitare nella barra della formula situata nella prima finestra disponibile all'avvio di Power BI Desktop.

Figura 45.

Le formule DAX vengono utilizzate in misure, colonne calcolate, tabelle e sicurezza a livello di riga.

- **Misure:** sono formule di calcolo dinamiche i cui risultati cambiano a seconda del contesto. Sono molto diverse dalle colonne calcolate in quanto, come detto, sono statiche e operano nel contesto del filtro corrente di un report; pertanto, le misure sono in continua evoluzione al variare del contesto del filtro. Le misure sono potenti strumenti analitici e, a causa del modo automatico in cui le misure funzionano con i contesti di filtro, sono sorprendentemente semplici da creare. Le misure vengono infatti utilizzate per la creazione di report che supportano la combinazione e il filtraggio dei dati dei modelli usando più attributi, ad esempio un report di Power BI oppure una tabella pivot o un grafico pivot di Excel. Le misure vengono create tramite la barra della formula DAX in "Progettazione modelli". Una formula in una misura può usare le funzioni di aggregazione standard create automaticamente tramite la funzionalità "Somma automatica", ad esempio COUNT o SUM, oppure può essere definita dall'utente tramite la barra della formula DAX. Quando si definisce una formula per una misura nella barra della formula, tramite una descrizione comando viene eseguito un'anteprima dei possibili risultati per il totale nel contesto corrente. Generalmente i risultati non vengono restituiti immediatamente in nessuna posizione. Il motivo per cui non è possibile visualizzare i risultati (filtrati) del calcolo è dovuto al fatto che il risultato di una misura non può essere determinato senza contesto. La valutazione di una misura richiede la presenza di un'applicazione client di creazione di report in grado di fornire il contesto necessario per recuperare i dati relativi a ogni cella e valutare l'espressione per ogni cella. Il client può essere una tabella pivot o un grafico pivot di Excel, un report di Power BI o un'espressione di tabella in una query DAX in SQL Server Management Studio (SSMS). Le misure possono distinguersi in implicite ed esplicite. Le aggregazioni implicite si verificano automaticamente sulle colonne con tipi di dati numerici. È generalmente considerata una buona pratica creare misure esplicite invece di misure implicite.

  Una misura esplicita presenta diversi vantaggi:

- Le misure possono essere costruite l'una sull'altra.
- Definiscono centralmente la formattazione dei numeri, creando coerenza.
- Incapsulano il codice, rendendo le modifiche logiche meno dispendiose in termini di tempo.

Le misure calcolate possono eseguire le seguenti operazioni:

- Possono essere assegnate a qualsiasi tabella.
- Interagiscono automaticamente con tutte le relazioni nel modello di dati, a differenza di quanto avviene sub colonne.
- Non sono materializzati in una colonna e pertanto non possono essere convalidati nella visualizzazione dati.

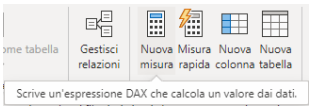


Figura 46.

- **Colonne calcolate:** consiste nell'aggiunta di una colonna a una tabella esistente in "Progettazione modelli" e per la quale viene creata una formula DAX che ne definisce i valori. Ciò permette di far sì che, per ogni riga di questa stessa colonna, vengono determinati i rispettivi valori non appena si immette la formula; questi ultimi vengono quindi archiviati nel modello di dati in memoria e sono ottenuti e visualizzati. I valori delle colonne vengono ricalcolati solo se la tabella, o una tabella correlata, viene elaborata (aggiornata) oppure se il modello viene salvato dalla memoria e quindi ricaricato, ad esempio quando si chiude e si riapre un file di Power BI Desktop. Le colonne calcolate vengono quindi archiviate nella tabella in cui sono assegnate e i valori rimangono i medesimi salvo aggiornamenti. Esistono molti casi d'uso per le colonne calcolate, ma i due più comuni sono i seguenti:
  - Attributi descrittivi
  - Colonne chiave concatenate

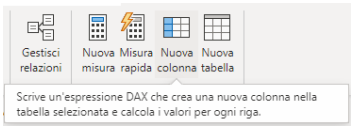


Figura 47.

- **Tabelle calcolate:** sono oggetti calcolati, basati su un'espressione di formula e derivate da tutte le altre tabelle nello stesso modello, o da alcune di esse. Invece di eseguire query e caricare i valori nelle colonne della nuova tabella da un'origine dati, una formula DAX definisce i valori della tabella. Le tabelle sono utili anche per la configurazione di un set di righe filtrato o di un subset o superset di colonne di altre tabelle esistenti. Ciò consente di mantenere intatta la tabella originale durante la creazione di varianti della tabella per supportare specifici scenari. Le tabelle misurano le relazioni con altre tabelle. Le colonne nella tabella calcolata contengono i tipi di dati, la formattazione e possono appartenere a una categoria di dati. È possibile assegnare un nome alle tabelle e rivelare o nasconderle esattamente come qualsiasi altra tabella. Le tabelle misurate vengono ricalcolate se una delle tabelle da cui vengono estratti i dati viene aggiornata.

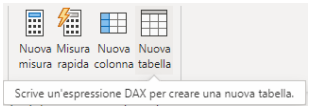


Figura 48.

- **Sicurezza a livello di riga:** in questo contesto, una formula DAX restituisce una condizione booleana VERO / FALSO, definendo le righe che possono essere restituite dai risultati di una query, in base ad un ruolo specifico. Quando si definisce la sicurezza a livello di riga pertanto, usando una formula DAX, si crea un set di righe permesse. Ciò non significa che viene negato l'accesso alle altre righe ma, semplicemente, che quest'ultime non vengono restituite come parte del set delle consentite. Altri ruoli, comunque, possono consentire l'accesso alle righe escluse dalla formula DAX. Se un utente è membro di un altro ruolo e la sicurezza a livello di riga di tale ruolo consente l'accesso a quel particolare set di righe, l'utente può visualizzare i dati per la riga. Le formule per la sicurezza a livello

di riga si applicano alle righe specificate e a quelle correlate. Quando una tabella ha più relazioni, i filtri applicano la sicurezza per la relazione attiva. Le formule per la sicurezza a livello di riga vengono intersecate con altre formule definite per le tabelle correlate.

2.2.1.8 Visualizzazione dati

Una volta che ci si è dedicati all'importazione dei dati e alla loro modellazione, secondo le proprie specifiche, è possibile realizzare delle loro visualizzazioni nel modo più efficiente ed efficace possibile. L'associazione infatti più comune con il software Power BI, per i consumatori, è proprio la capacità di creare visualizzazioni di dati di grande impatto, anche grazie alle molte opzioni disponibili.

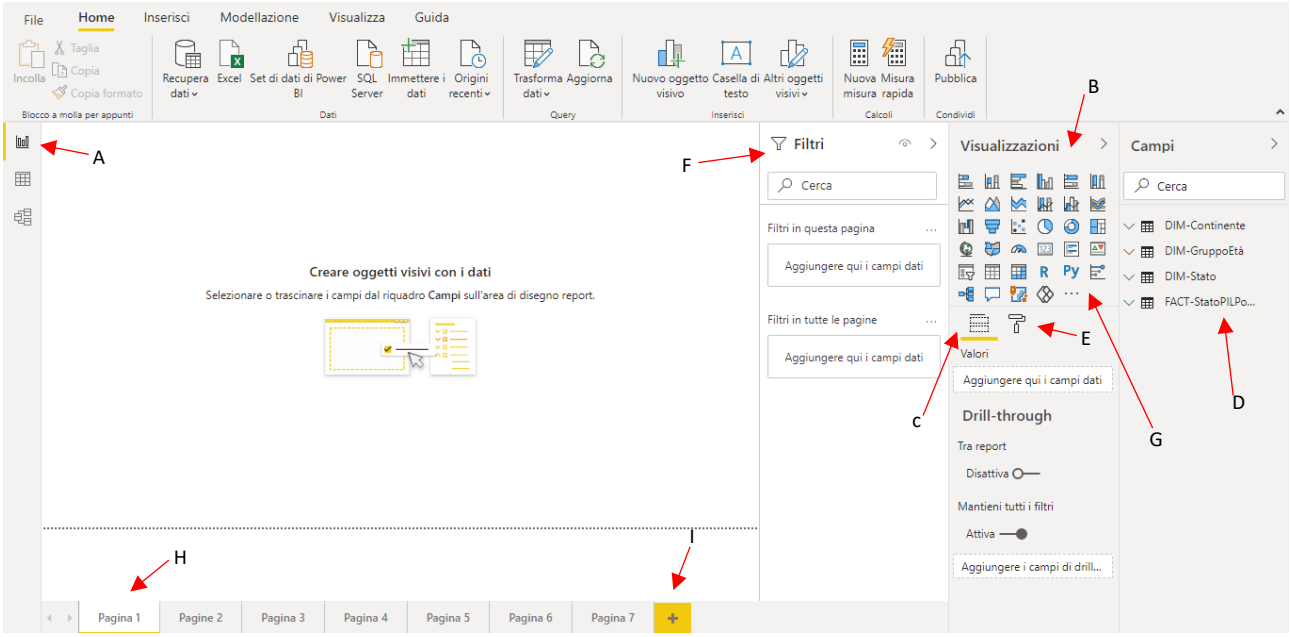


Figura 49.

- **A = Visualizzazione report:** questo è il pulsante che ci posizionerà nell'area di disegno del report e ci consentirà di creare immagini.
- **B = Area degli elementi visivi:** qui è si può scegliere quale elemento visivo si vuole utilizzare.
- **C = Area dei valori:** quest'area cambierà a seconda della visuale, ma è dove posizioneremo i campi che useremo all'interno della visuale selezionata.

- **D = Riquadro dei campi:** questi sono tutti i campi disponibili che dobbiamo scegliere da aggiungere alle nostre immagini.
- **E =Area formato:** qui è dove possiamo decidere su molte cose specifiche dell'intera pagina del report o dell'oggetto visivo selezionato, come la dimensione del testo, lo stile del carattere, i titoli e così via.
- **F = Area filtri:** qui è dove possiamo applicare filtri di vari ambiti:
    - **Filtri a livello di pagina:** qualsiasi filtro applicato qui influenzerà ogni singolo elemento visivo sulla pagina selezionata.
    - **Filtri drill-through:** questa opzione consente agli utenti di passare un valore di filtro da una pagina del report diversa a questa.
    - **Filtri a livello di report:** i filtri applicati qui avranno effetto su ogni singolo oggetto visivo per l'intero report di Power BI.
    - **Filtri a livello visivo:** questa categoria appare solo quando si seleziona un oggetto visivo e i filtri applicati influenzeranno solo l'oggetto visivo selezionato stesso.
- **G = Elementi visivi personalizzati:** selezionando questo pulsante, verrà visualizzato un menu che ha accesso a tutti gli elementi visivi personalizzati da Microsoft Store. È quindi possibile aggiungere qualsiasi elemento visivo che si desidera all'area degli elementi visivi.
- **H = Pagina del rapporto:** qui è dove si può selezionare la pagina del rapporto con cui si desidera lavorare. Ogni pagina ha un'area di lavoro limitata in cui è possibile usare gli oggetti visivi, quindi è comune avere più di una pagina in un report di Power BI.
- **I = Aggiungi pagina report:** selezionando questo simbolo, è possibile aggiungere una nuova pagina report per aggiungere più aree in cui è possibile aggiungere immagini.

È importante notare che quando si lavora con i filtri a livello di visualizzazione, l'area “campi” e l'area “formato” riguardano l’elemento visivo specifico.

Filtrare i dati che gli utenti vedranno all'interno di un report di Power BI, è il modo più efficace per rispondere a domande molto specifiche su quei stessi dati e ci sono molti modi per farlo. Non a caso, una delle migliori caratteristiche di Power BI è proprio la sua capacità di consentire agli utenti di interagire con un oggetto visivo, il quale applicherà, di conseguenza, il filtro appena selezionato al resto degli oggetti visivi in quella stessa pagina: questo è noto come filtro interattivo. È una caratteristica estremamente significativa in quanto fornisce un maggiore controllo direttamente nelle mani degli utenti, che possono decidere come filtrare le immagini. Insieme a questa funzionalità si ha la possibilità anche di aggiungere forme di filtro più esplicite, utilizzando l’apposito oggetto visivo: è in questa sezione che si potrà scegliere un campo più specifico dai dati, in relazione all’interesse degli utenti finali che li vorranno manipolare per vedere quei dati in vari stati diversi.

Power BI offre un’ampia gamma di visualizzazioni, differenti e utili a seconda dei dati a disposizione e di ciò che si vuole venga mostrato.

- **Filtro dei dati:** questo oggetto visivo, anche chiamato Slicer, consente l'assegnazione di uno o più campi e, a seconda del tipo di dati di quel campo, si hanno diverse opzioni di presentazione.

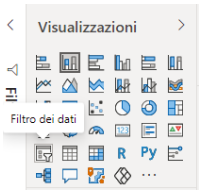


Figura 50.

Per questa visualizzazione, si hanno a disposizione diversi alternative:

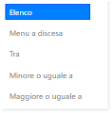


Figura 51.

- **Elenco:** questa opzione è disponibile indipendentemente dal campo selezionato. È un elenco distinto di valori dal campo selezionato. È meglio usarlo quando c'è una piccola quantità di opzioni tra cui scegliere.

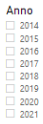


Figura 52.

Figura 53.

- **Menu a discesa:** fornisce all'utente un menu a discesa che conterrà un elenco distinto di valori dal campo selezionato. È molto simile all'opzione "Elenco", ma le scelte sono nascoste finché non si preme l'opzione a discesa. È pensato per un set di valori ancora più piccolo, in modo che gli utenti non debbano scorrere centinaia di scelte.

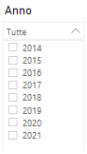


Figura 54.

- **Tra:** questa scelta si presenterà solo per i campi che sono di un tipo di dati numerico includendo pertanto anche le date. Consente agli utenti di specificare un intervallo di valori da sfruttare come filtro utilizzando una barra di scorrimento.

Figura 55.

- **Minore o uguale a:** molto simile all'opzione "Tra", ma la scala mobile può essere regolata solo dal lato sinistro.

Figura 56.

- **Maggiore o uguale a:** è uguale all'opzione precedente, tranne per il fatto che è possibile regolare la scala di scorrimento solo dal lato destro.

Figura 57.

Si hanno molte opzioni all'interno di Power BI per rappresentare visivamente i dati ma, a volte, si potrebbero voler vedere e confrontare dati dettagliati e valori esatti. In questi scenari, l'utilizzo dell'oggetto visivo “Tabella” o “Matrice” finisce per essere la scelta migliore. Quando si utilizza uno di questi due elementi visivi, è importante sfruttare l'area “Formato” per garantire che gli utenti possano interpretare facilmente i dati dettagliati presentati. Uno dei modi migliori per attirare l'attenzione sui valori di importanza con questi elementi visivi è usare la formattazione condizionale.

- **Tabella:** l'oggetto visivo della tabella è perfetto per esaminare molti valori (misure) per una categoria.

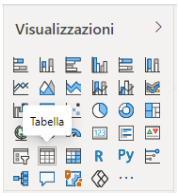


Figura 58.

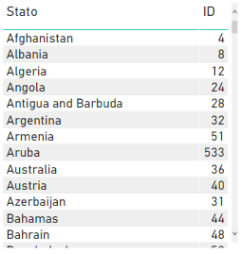


Figura 59.

Per far brillare davvero la tavola, si può anche sfruttare l'opzione di formattazione condizionale che è a disposizione.

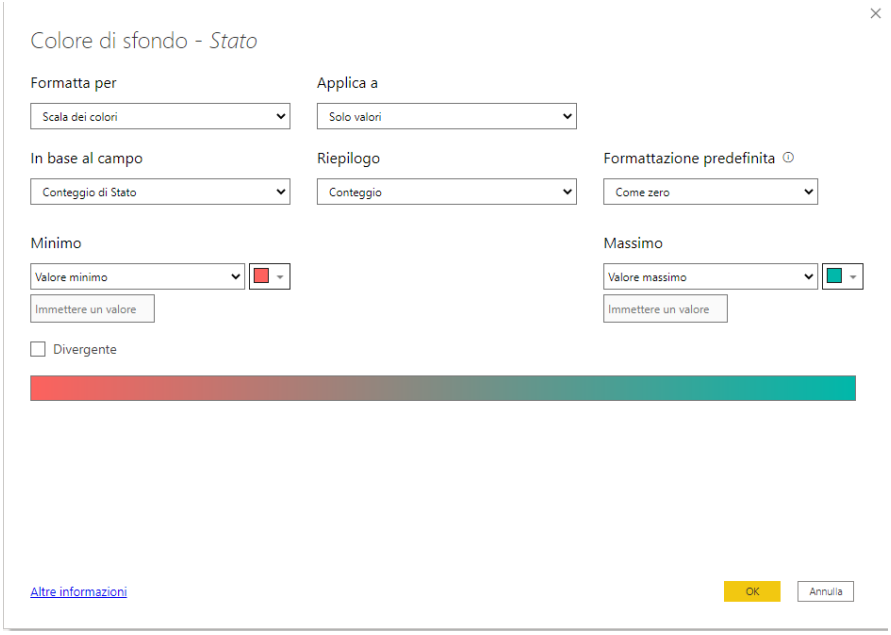


Figura 60.

- **Matrice:** laddove una tabella fa un ottimo lavoro nel consentire agli utenti di consumare una miriade di dati dettagliati su una singola categoria, l'elemento visivo “Matrice” può farlo per più di una categoria. Questo oggetto visivo infatti consente agli utenti non solo di selezionare una categoria per le righe, ma anche che si possa selezionare un campo per popolare le colonne, il che consente di vedere dati dettagliati in una sezione trasversale per due categorie. La formattazione condizionale è disponibile anche per l'uso all'interno di questo oggetto visivo ed è incorporata nello stesso modo in cui è stato eseguito nell'esempio precedente. Oltre alla formattazione condizionale, la matrice può sfruttare le gerarchie stabilite per fornire agli utenti la capacità di eseguire il drill-down in dati più granulari.

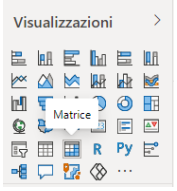


Figura 61.

Continente	2016	2017	2018	2019	2020	2021	Totale
AF	1.137.652.652,00	1.167.804.103,00	1.198.490.611,00	1.229.607.316,00	1.261.077.744,00	1.292.857.005,00	**9.474.846.732,00**
AS	4.568.156.236,00	4.610.877.245,00	4.652.520.169,00	4.692.899.841,00	4.731.881.729,00	4.762.592.990,00	**37.023.505.622,00**
EU	859.265.951,00	862.067.756,00	864.647.645,00	866.836.585,00	868.519.924,00	869.638.575,00	**6.900.791.670,00**
NA	560.240.682,00	564.996.363,00	569.649.544,00	574.237.002,00	578.786.079,00	583.307.294,00	**4.536.914.496,00**
OC	39.372.123,00	39.935.188,00	40.491.798,00	41.041.237,00	41.583.166,00	42.117.513,00	**321.574.658,00**
SA	415.894.121,00	419.625.657,00	423.294.926,00	426.905.230,00	402.021.664,00	405.245.213,00	**3.313.321.890,00**
Totale	**7.580.581.765,00**	**7.665.306.312,00**	**7.749.094.693,00**	**7.831.527.211,00**	**7.883.870.306,00**	**7.955.758.590,00**	**61.570.955.068,00**

Figura 62.

Se gli oggetti visivi Tabella e Matrice consentono un'analisi dettagliata di più misure in una categoria, gli oggetti visivi seguenti sono i migliori per visualizzare i valori dei dati tra le categorie. Nelle prossime immagini, visualizzeremo barre, colonne e altri elementi visivi, che saranno proporzionali al valore dei dati. Questi elementi visivi hanno una quantità di dati molto meno dettagliata, ma è molto facile distinguere le differenze dei valori all'interno della categoria scelta. Tutti gli elementi visivi consentono il filtraggio interattivo e l'uso di drilldown.

- **Grafici a barre e a colonne:** sono molto simili nella configurazione e nel modo in cui visualizzano i dati. L'unica differenza è l'orientamento: il grafico a barre utilizza barre rettangolari orizzontalmente dove la lunghezza della barra è proporzionale ai dati, mentre l'istogramma mostra le barre verticalmente; entrambi vengono però utilizzati per confrontare due o più valori. Entrambe queste visualizzazioni hanno inoltre anche diversi formati.

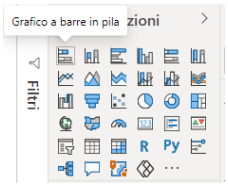


Figura 63.

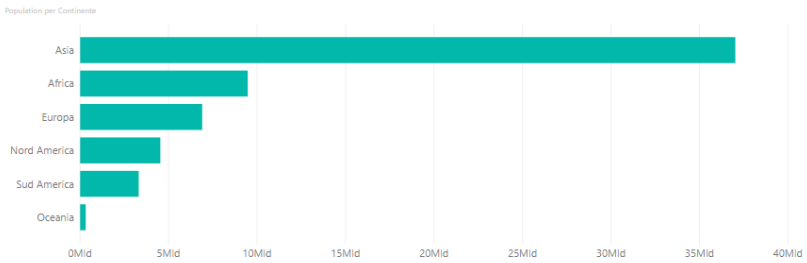


Figura 64.

- **Grafici a torta e ad anello:** entrambi gli oggetti hanno lo scopo di visualizzare una particolare sezione nel suo insieme piuttosto, invece, che confrontare i singoli valori tra loro. La principale differenza tra i due grafici, è che il grafico ad anello ha un buco nel mezzo che potrebbe consentire una sorta di etichetta. Entrambe queste immagini possono essere molto efficaci nel consentire il filtraggio interattivo, ma se ci sono troppe categorie può diventare difficile da leggere ed interpretare. Per rendere ancora più facile la lettura, è possibile ricorrere alle etichette di dettaglio. Questa opzione è molto simile alle etichette di categoria in quanto si ha modo di visualizzare i dati di ciascuno dei quadranti senza dover utilizzare i suggerimenti. Tuttavia, la presenza di tanti elementi come etichette può causare a sua volta una non trascurabile confusione.

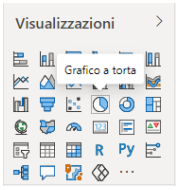


Figura 65.

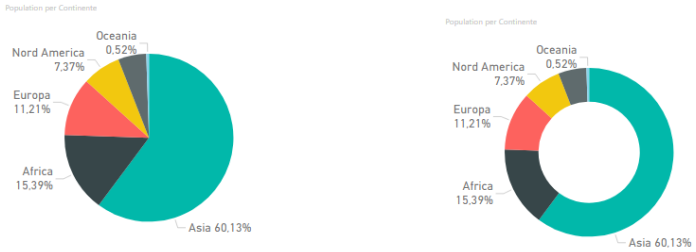


Figura 66.

- **Mappa ad albero:** a risultare conforme per la visualizzazione delle gerarchie è la visuale della mappa ad albero. Lo fa annidando i dati in rettangoli, che sono rappresentati dal colore, comunemente noto come "ramo". Se si aggiunge una categoria dalla sezione “dettagli” dell'oggetto visivo, si noteranno ulteriori rettangoli più piccoli all'interno dei "rami" e questi sono noti come "foglie", da cui il nome “Treemap”.

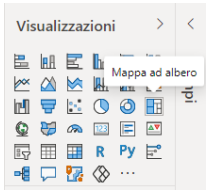


Figura 67.

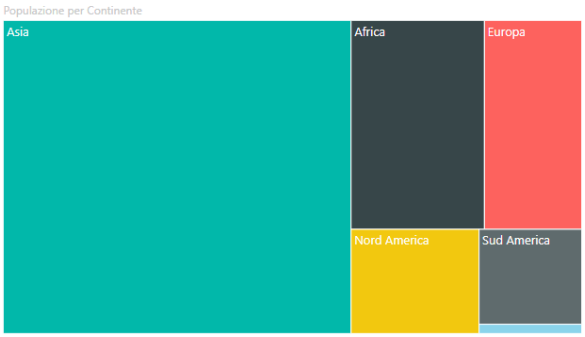


Figura 68.

- **Il grafico a dispersione:** a volte indicato come grafico a bolle, questa visual consente di mostrare le relazioni tra due o tre valori numerici. Ci viene data l'opportunità di posizionare valori per gli assi x e y ma la differenza in questo elemento visivo è che si può aggiungere un terzo valore per la dimensione, ed è da qui che viene proprio il nome di grafico a bolle.

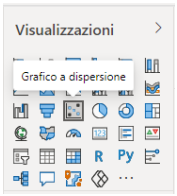


Figura 69.

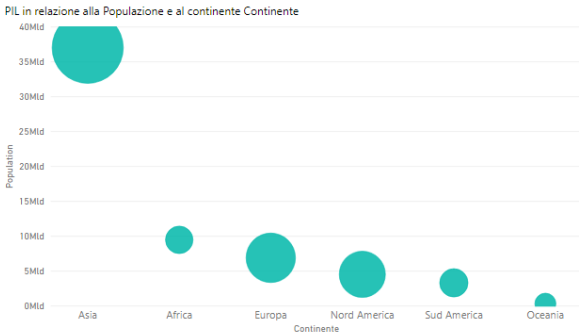


Figura 70.

Quando si usa il termine Trend Data, si sta parlando di visualizzare e confrontare i valori nel tempo. Power BI offre molte opzioni in questa categoria, ognuna con il proprio focus. L'idea per ciascuna delle immagini, tuttavia, è di attirare l'attenzione sul valore totale per un periodo di tempo.

- **Grafici a linee e ad area:** è la più semplice delle nostre opzioni quando si tratta di esaminare i dati nel tempo. Il grafico ad area e il grafico ad area in pila si basano sul grafico a linee; la differenza è che l'area tra l'asse e la linea viene riempita di colori per mostrare il volume.

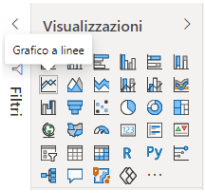


Figura 71.

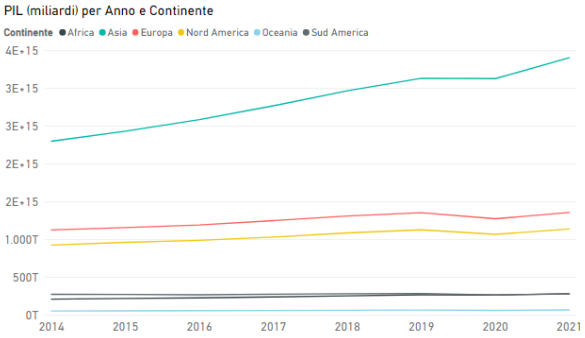


Figura 72.

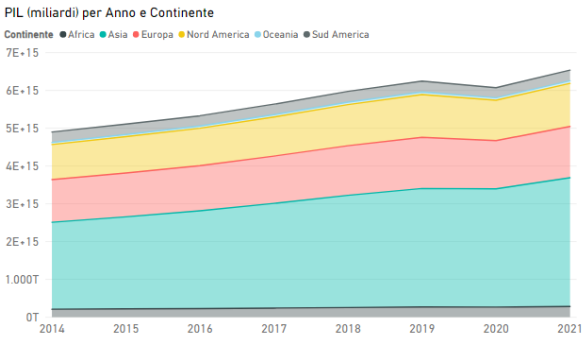


Figura 73.

- **Grafici combinati:** combinano il grafico a linee e l'istogramma insieme in un unico oggetto visivo. Gli utenti possono scegliere di avere il formato Stack Column o il

formato Clustered Column. Combinando questi due elementi visivi insieme, è possibile fare un confronto molto rapido dei dati. Il vantaggio principale di questo tipo di grafico è che si possono avere uno o due assi Y: ciò significa che si potranno visualizzare due misure che avrebbero lo stesso asse Y.

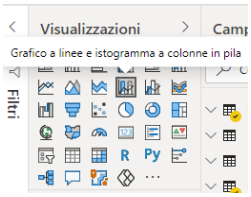


Figura 74.

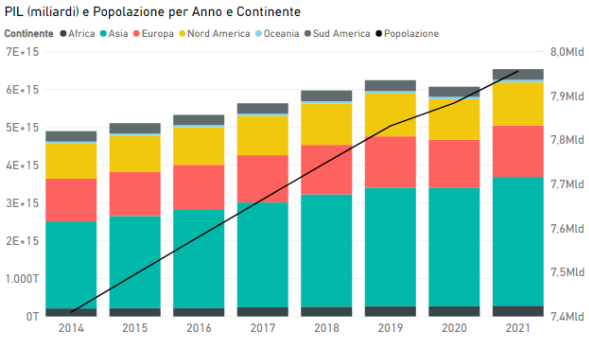


Figura 75.

- **Il grafico a nastro:** non è diverso dagli altri elementi visivi in quanto è utile per visualizzare i dati nel tempo. Ciò che rende i grafici a nastro efficaci, tuttavia, è la loro capacità di mostrare il cambio di grado; l'intervallo o il valore più alto viene sempre visualizzato in alto per ciascuno dei periodi di tempo. Il grafico ha anche un'attrattiva visiva unica, diversa dagli altri elementi visivi.

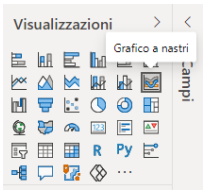


Figura 76.

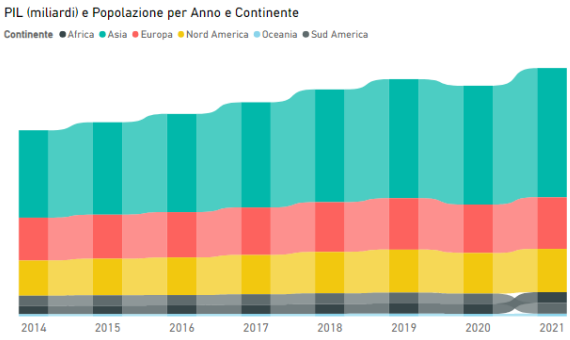


Figura 77.

- **Grafico a cascata:** è molto utile per comprendere i cambiamenti che si verificano da un valore iniziale. Visualizza un totale parziale in relazione ai valori aggiunti o sottratti. Compilando un campo nell'opzione “ripartizione” dell'oggetto visivo, si può vedere se ha avuto un impatto positivo o negativo da valore a valore.

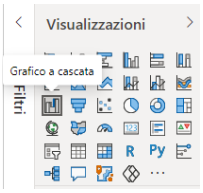


Figura 78.

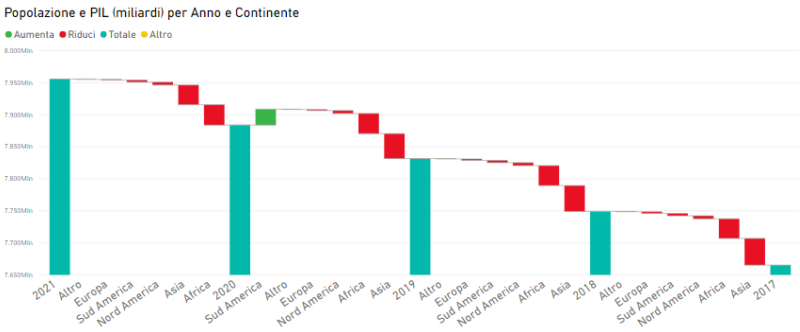


Figura 79.

- **Grafico a imbuto:** consente agli utenti di vedere la differenza percentuale tra i valori. Normalmente, il valore più alto è in alto e il più basso in basso, il che dà l'aspetto di un imbuto. Ogni fase del funnel racconta la differenza percentuale tra se stessa e la fase precedente, oltre che rispetto alla fase più alta. Con questo tipo di design ha senso che il grafico a imbuto sia molto efficace quando si visualizza

un processo lineare con almeno tre o quattro fasi. Il nostro set di dati non ha un processo con più fasi, ma possiamo comunque creare qualcosa che ci dia valore.

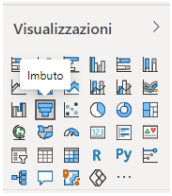


Figura 80.

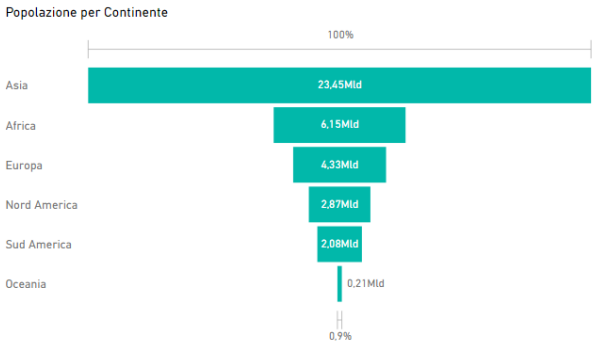


Figura 81.

Key Performance Indicator (KPI), è un valore misurabile che dimostra quanto un'azienda stia raggiungendo un determinato obiettivo. Con Power BI, si hanno un paio di opzioni per misurare i progressi compiuti verso un obiettivo per i processi operativi. La forza di una visualizzazione KPI risiede nella sua semplicità. Visualizza un singolo valore e il suo progresso verso un obiettivo specifico.

.

- **Indicatore:** mostra un singolo valore all'interno di un arco circolare e il suo progresso verso un obiettivo o un valore di destinazione specificato. Il valore target è rappresentato da una linea all'interno dell'arco.

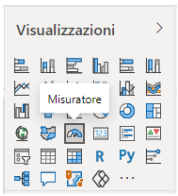


Figura 82.

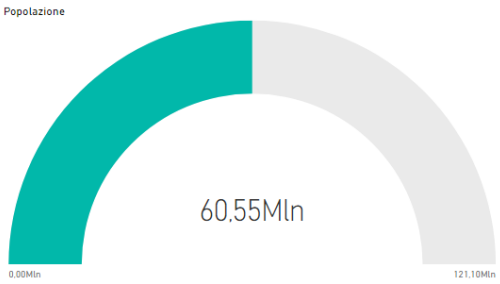


Figura 83.

- **L'oggetto visivo KPI:** se l' "indicatore" usa l'arco circolare per mostrare lo stato di avanzamento corrente, l'oggetto visivo KPI ha un approccio più esplicito e mostra solo il valore in testo normale, insieme all'obiettivo. Gli unici elementi visivi reali che sono in gioco con questa immagine si verificano quando il valore dell'indicatore è inferiore all'obiettivo e il testo è mostrato in rosso, e quando ha superato l'obiettivo e il testo è in verde. Questa è sicuramente una delle immagini più dirette ed esemplifica perfettamente ciò che si attende da un KPI.

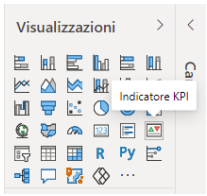


Figura 84.

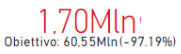


Figura 85.

Uno dei modi più interessanti per visualizzare i dati in Power BI è attraverso le varie opzioni di mappa a disposizione. Tutte le mappe hanno lo scopo di illustrare i dati in relazione alle località in tutto il mondo, ma ci sono alcune piccole differenze tra ciascuna di esse. Tutte le mappe, eccetto la Shape Map, hanno l'opzione di fornire le coordinate di latitudine e longitudine, che saranno il modo migliore per garantire che venga visualizzata la posizione appropriata. Il motivo è perché le informazioni che si forniscono alla visuale verranno inviate a Bing Maps per verificare il posizionamento

sulla mappa. Se non si forniscono dettagli sufficienti, Bing potrebbe non restituire i risultati desiderati. Ad esempio, se si dovesse fornire alla visuale della mappa un campo contenente solo il nome della città, ciò potrebbe creare confusione perché potrebbero esserci più città con quel nome. In questi scenari, sarebbe meglio fornire una sorta di gerarchia geografica per dare una migliore definizione oppure, creare nuove colonne con informazioni più dettagliate. Power BI ha anche una funzionalità incorporata quando si tratta di dati geografici che consente agli utenti di identificare il tipo di dati forniti: questa è chiamata “Categoria di dati”.

- **Mappa:** questo oggetto visivo è noto anche come “Bubble Map” perché traccia i punti di dati con cerchi che possono essere impostati per modificare le dimensioni in base a una misura fornita. Con questo oggetto visivo, se si hanno le coordinate di latitudine e longitudine nel proprio set di dati, non è necessario inviare nulla a Bing Maps.

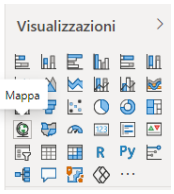


Figura 86.

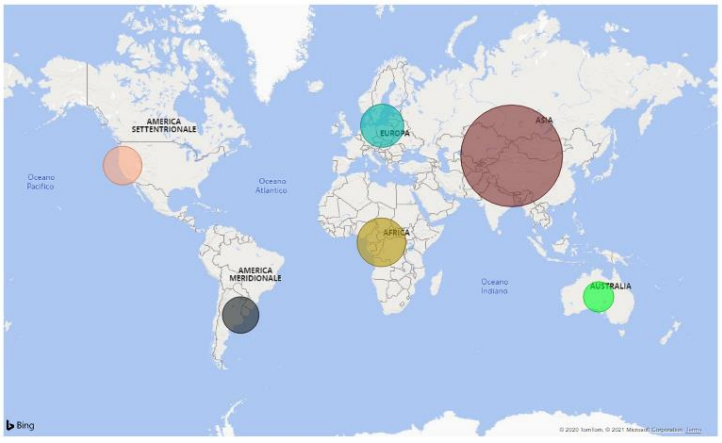


Figura 87.

- **Mappa colorata:** a differenza dell'elemento visivo “mappa” tradizionale, che utilizza una bolla per indicare le posizioni, questo elemento visivo usa l'ombreggiatura per visualizzare i dati geografici. Quindi, più chiara appare

un'area, minore è il valore rappresentativo. Per questo elemento visivo, è consigliabile rifarsi all'area formato e dettare la gamma di colori per l'ombreggiatura in modo che appaia più evidente.

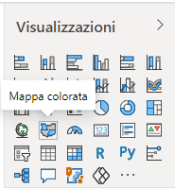


Figura 88.

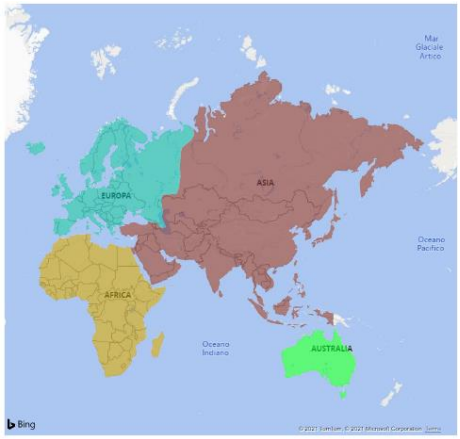


Figura 89.

Sono stati mostrati molti elementi visivi diversi e come funzionano con tipi specifici di dati. Nonostante questi, si ha comunque la possibilità di selezionare l'intera raccolta di elementi visivi personalizzati disponibili forniti sempre dalla stessa Microsoft. È importante sapere che quando si seleziona un oggetto visivo personalizzato, questo viene salvato come parte del file di report di Power BI e non rimane all'interno dell'applicazione.

2.2.2 Power BI Service

Power BI, come detto, è una raccolta di servizi software, app e connettori che interagiscono per semplificare la creazione, la condivisione e l'utilizzo di informazioni dettagliate aziendali nel modo più efficace per l'utente e l'azienda. Il servizio Microsoft Power BI, noto anche come Power BI Online, è la parte SaaS (Software as a Service) di Power BI. Dopo aver creato report e dashboard, è possibile condividerli in modo che gli

utenti finali nel servizio Power BI e nei dispositivi mobili possono visualizzarli e interagire con essi. La possibilità di controllare la modalità di condivisione del lavoro è una delle funzionalità più importanti del servizio Power BI. È possibile creare aree di lavoro per collaborare con i colleghi, è quindi possibile aggregarli e distribuirli come app. Si possono anche condividere i set di dati, in modo che possono essere usati da altri come base per i propri report.

Nel servizio Power BI, infatti, i dashboard aiutano a tenere sotto controllo l'andamento dell'attività. I dashboard mostrano riquadri che è possibile selezionare per aprire i report e ottenere informazioni ancora più dettagliate. I dashboard e i report si connettono ai set di dati e pertanto riuniscono tutti i dati rilevanti insieme in un'unica posizione. In un flusso di lavoro tipico di Power BI si crea un report in Power BI Desktop per poi pubblicarlo nel servizio Power BI ma è anche possibile creare report di Power BI direttamente nel servizio Power BI.

2.2.2.1 Importazione e caratteristiche principali

Esistono numerosi modi per pubblicare un report nel servizio PowerBI, ma il modo più semplice è usare il desktop di Power BI. È possibile fare ciò anche mediante l'applicazione desktop.

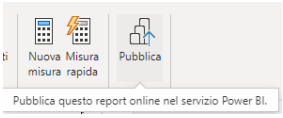


Figura 90.

È richiesto a questo punto in quale area di lavoro si desidera eseguire la condivisione. Un'area di lavoro è un'area del servizio Power BI che è molto simile a una cartella, in cui è possibile raggruppare report, set di dati e dashboard. È prevista anche l'assegnazione di zone e funzioni protette nell'area. Ancora più importante, consente come detto lo sviluppo in team di una soluzione Power BI, in cui è possibile avere più autori su una soluzione.

Se si apre il rapporto, PowerBI.com si avvierà e mostrerà lo stesso rapporto, che si stava visualizzando sul desktop, in un browser web. PowerBI.com dispone di quattro aree chiave con cui è possibile interagire:

- **Set di dati:** si tratta dei dati non elaborati che sono stati creati nel desktop di Power BI. È possibile creare anche un nuovo set di dati.
- **Cartelle di lavoro:** queste cartelle di lavoro possono essere usate come set di dati o possono formare parti della cartella di lavoro che possono essere aggiunte a un dashboard.
- **Report:** si riferisce a ciò che è stato creato nel desktop di Power BI. Questi possono essere esplorati, modificati o scaricati in questa stessa sezione.
- **Dashboard:** finalizzato come raccolta degli elementi migliori dei report in un set unificato di dashboard. Questi sono la prima finestra con cui interagirà la maggior parte degli utenti occasionali.

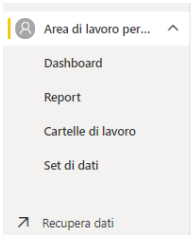


Figura 91.

2.2.2.2 Creazione dashboard

Su ciascuno dei grafici, riquadri e altri elementi visivi, vi è un'icona a forma di puntina in alto a destra di quell'oggetto. Una volta selezionato, viene richiesto a quale dashboard si desidera aggiungere il report. È possibile, a quel punto, selezionare una nuova dashboard per creare o scegliere una dashboard esistente a cui aggiungere l'elemento.

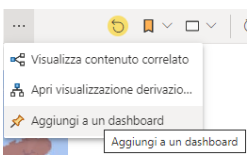


Figura 92.

Figura 93.

Una volta completata la dashboard, si possono porre domande sui propri dati. Proprio sopra i dati dei dashboard, è presente l'area in cui poter porre una domanda sui dati. Ad esempio, si può porre la domanda "Mostrami gli archivi totali per stato" e Power BI produrrà una risposta geografica a quella domanda. Se si preferisce vedere la risposta come un grafico a barre invece che come una mappa, è possibile richiederla esplicitamente come un elemento grafico. Se piace la risposta che viene restituita, si può aggiungere l'elemento visivo nell'angolo in alto a destra per appuntare l'elemento del rapporto a una dashboard. È inoltre possibile espandere i filtri e le visualizzazioni per essere molto precisi con l'elemento visivo in questione. Sebbene Power BI sia ottimo per rispondere alle domande con i filtri, a volte necessita di una regolazione fine.

2.2.2.3 Condivisione

La condivisione in Power BI è abbastanza semplice, ma vi sono diverse alternative a riguardo. Si può infatti condividere una versione di sola visualizzazione di un rapporto o dashboard con cui gli utenti potrebbero interagire o, se l'obiettivo è consentire agli utenti di modificare anche il rapporto, di utilizzare gli spazi di lavoro. Infine, se si desidera raggruppare in modo logico report e dashboard e avere la possibilità di avere il controllo preciso su quali report possono essere visualizzati per impostazione predefinita, di prendere in considerazione l'utilizzo delle app Power BI. Il modo più semplice per condividere un dashboard o un report è dall'apposito pulsante, semplicemente inserendo l'indirizzo email dell'utente con cui si desidera condividere e il tipo di accesso che si vuole concedere. Anche se non si vuole consentire loro di modificare il rapporto o

la dashboard, questi saranno comunque in grado di visualizzare e ricondividere il rapporto. In qualsiasi momento, si può anche vedere quali risorse sono condivise.

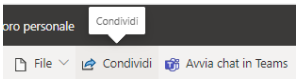


Figura 94.

2.2.2.4 Pianificazione degli aggiornamenti dei dati

Il servizio Power BI ha la possibilità di aggiornare i set di dati. Se tutti i dati risiedono nel cloud, l'aggiornamento è molto semplice. Tuttavia, se si dispone di alcuni dati o file in locale, è necessario installare il gateway locale.

I gateway locali possono essere usati su più servizi cloud, come Power BI, PowerApps, App per la logica e Microsoft Flow. È possibile scaricare il gateway gratuito dall'icona di download su PowerBI.com una volta effettuato l'accesso. La prima domanda che verrà posta durante l'installazione è se si desidera installare il gateway dati in modalità personale o in locale modalità gateway. La differenza principale tra il gateway dati locale e il gateway dati locale in modalità personale è che la modalità personale viene eseguita come un'applicazione rispetto a un servizio Windows. Installando in modalità personale pertanto, si rischia che dati diventino obsoleti se l'applicazione non è aperta all'avvio del PC. È utile per quegli utenti che potrebbero non avere accesso amministrativo alla propria macchina o per gli utenti che desiderano aggiornamenti dei dati più semplici.

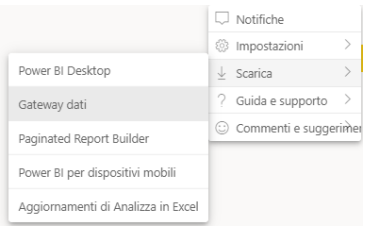


Figura 95.

CAPITOLO III

Caso studio

3.1 Descrizione

In corrispondenza del termine del percorso tirocinante intrapreso all'interno della società ospitante, una volta pertanto giunti alla conclusione della fase di preparazione, formazione e studio delle differenti tematiche concernenti la Business Intelligence, mi è stata assegnata la risoluzione della simulazione di un attinente progetto, riguardante una società (denominata semplicemente Alfa), al fine che ne implementassi una soluzione all'avanguardia, in linea con le tematiche affrontate nell'arco temporale di svolgimento del rapporto.

La società Alfa, sulla quale si è incentrato il progetto, è stata rappresentata come una medio/piccola realtà, inserita all'interno del settore della consulenza informatica. L'operatività svolta dall'Alfa si colloca in un tessuto territoriale con spiccata vocazione industriale, principalmente limitrofo alle sue sedi piemontesi, con clienti eterogenei, operanti in differenti contesti e ambiti e di diversa dimensione, dalla piccola realtà locale fino al cliente di rilevante spessore. I servizi offerti prevedono principalmente l'assistenza nell' ICT, nello sviluppo mobile e web e, inoltre, nella gestione dei big data. Le dimensioni effettive corrispondono ad una realtà che occupa circa un centinaio di dipendenti, con una crescita costante negli ultimi anni, parallela all'incremento del fatturato sviluppato. L'attenzione e l'investimento sui giovani è significativo; la Alfa offre infatti all'interno del suo sito diverse sezioni dedicate all'inserimento e alla formazione di nuove figure.

Per quanto concerne il progetto invece, questo è consistito nella disposizione di un set di dati, estrapolati ed elaborati dal servizio di Google Analytics, relativi alle statistiche sui dati di navigazione degli utenti, sulle pagine/sezioni visitate quindi e sulle visualizzazioni avvenute in un determinato periodo di tempo, riguardo gli accessi al sito web societario [115, 116].

Una volta ricevuti i dati, sono state applicate le nozioni acquisite durante il periodo di apprendimento sulle caratteristiche della Business Intelligence. A queste, sono state accompagnate le conoscenze tecniche di uno strumento che ne permette l'esplicazione: il software Power BI. Scopo finale era infatti la costituzione di una dashboard, un mezzo capace di proporre, in una visione di alto livello, le relazioni e le informazioni ricavate dal relativo set di dati; un prodotto quindi in grado di fornire un supporto decisionale alle diverse sezioni amministrative della società Alfa.

Il processo di risoluzione è iniziato con la ricezione ed il caricamento del set di dati sulla componente locale Power BI Desktop. Conseguentemente, è proseguito con l'anticipo, seppur momentaneo, di quanto occorresse in realtà realizzare come step conclusivo del processo di manipolazione e pulizia dei dati; le diverse tabelle, costituenti il database, sono state infatti riportate in report e differenti grafici, al solo scopo di visionare ed intrepretare, seppur in linee generali, le varie funzioni delle relative colonne.

Visionate complessivamente le correlazioni e le caratteristiche generiche del set, si è potuto iniziare l'effettivo processo di trasformazione, con il quale si ha avuto modo di organizzare, filtrare e ripulire i dati a disposizione. Questa fase è stata realizzata in un ambiente apposito di Power BI Desktop chiamato "Query editor".

Terminata questa fase, si è proseguito con la sezione riguardante la creazione delle connessioni tra le diverse tabelle, la realizzazione delle misure e la definizione di report e grafici.

3.2 Data set

I dati dai quali è stato possibile estrarre le informazioni, tali da poter delineare un ausilio, un supporto alle scelte decisionali dell'azienda Alfa e attinenti all'efficienza del sito web (in termini di visualizzazioni, interazioni, tempi di permanenza e allocazione geografia di accesso degli utenti) sono stati elaborati ed estrapolati dal servizio di Google Analytics. Quest'ultimo, come intuibile, altro non corrisponde che ad un prodotto offerto da Google, implementato e adoperato con l'intento di fornire statistiche e strumenti

analitici indirizzati al monitoraggio delle attività avvenute sul proprio sito web; consente pertanto di valutare il rendimento a livello di marketing, prodotti e contenuti oltre che consentire l'accesso alle funzioni di machine learning. Google Analytics è un servizio che offre la possibilità di reperire informazioni sugli utenti e di conseguenza permetterne di delinearne diverse tipologie, in modo che possano considerarsi nuove soluzioni o il focalizzarsi su alcune di queste stesse tipologie, in base alle prospettive di breve o lunga durata del proprio business.

Il data set costruito quindi, è stato strutturato in funzione degli aspetti peculiari del traffico di rete avvenuto sulle pagine web dell'azienda. Le informazioni di maggiore rilevanza hanno riguardato, con le rispettive tabelle e interconnessioni fra queste stesse, tematiche quali il tracciamento delle pagine, la geografia degli accessi, il numero di utenti coinvolti e la loro tipologia, le loro interazioni all'interno del sito e le sessioni avviate.

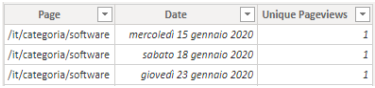
- **Page tracking:** tabella di raccolta delle informazioni attinenti al tracciamento per ogni singola pagina del sito nei diversi giorni, distinta per rispettivo indirizzo.

Date	Unique Pageviews	Time on Page	Avg. Time on Page	Unique Events	Page
sabato 9 novembre 2019	2	00:00:00	00:00:00	2	/hib
martedì 10 settembre 2019	1	00:00:00	00:00:00	1	/hib
mercoledì 12 febbraio 2020	1	00:00:00	00:00:00	1	/crf/kpr

Figura 96.

  - **Date:** colonna riguardante le date prese in considerazione.
  - **Unique Pageviews:** aggrega le visualizzazioni della pagina generate dallo stesso utente durante la stessa sessione. Una visualizzazione di pagina unica (unique pageviews per l'appunto) rappresenta il numero di sessioni durante le quali quella pagina è stata visualizzata una o più volte.
  - **Time on Page:** tempo totale trascorso sulla pagina.
  - **Avg. Time on Page:** tempo medio trascorso sulla pagina.
  - **Unique Events:** aggrega gli eventi scaturiti sulla pagina dallo stesso utente durante la stessa sessione.
  - **Page:** indirizzo o URL della pagina.

- **Page Tracking Blog:** analogamente alla tabella denominata Page Tracking, questa presenta la medesima funzione, ma riguardante le sole pagine attinente alla sezione “Blog” del sito web.

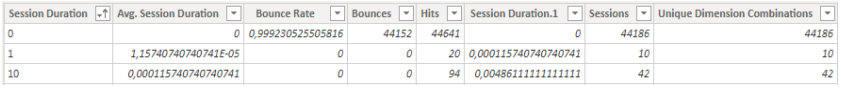


Page	Date	Unique Pageviews
/it/categoria/software	mercoledì 15 gennaio 2020	1
/it/categoria/software	sabato 18 gennaio 2020	1
/it/categoria/software	giovedì 23 gennaio 2020	1

Figura 97.

  - **Page:** indirizzo o URL della pagina
  - **Date:** colonna riguardante le date prese in considerazione.
  - **Unique Pageviews:** aggrega le visualizzazioni della pagina generate dallo stesso utente durante la stessa sessione. Una visualizzazione di pagina unica (unique pageviews per l’appunto) rappresenta il numero di sessioni durante le quali quella pagina è stata visualizzata una o più volte.

- **Session:** tabella che racchiude i dettagli sulle sessioni avviate da tutti gli utenti che hanno avuto accesso al sito. Da queste, strumenti lato serve, si ricavano informazioni sul tempo totale di navigazione e non solo. Se però un utente è inattivo sul sito per 30 minuti o più, qualsiasi attività futura viene attribuita a una nuova sessione. Gli utenti che abbandonano il sito e ritornano entro 30 minuti vengono conteggiati come parte della sessione originale. La prima sessione in assoluto di un utente è considerata una sessione aggiuntiva e un utente aggiuntivo. Eventuali sessioni future dello stesso utente vengono conteggiate come sessioni aggiuntive, ma non come utenti aggiuntivi. Le sessioni vengono incrementate con il primo hit di una sessione, mentre le entrate vengono incrementate con il primo hit di visualizzazione di pagina di una sessione. Se il primo hit della sessione non é una visualizzazione di pagina, si potrà rilevare una differenza tra il numero di sessioni e il numero di ingressi [117, 118].



Session Duration	Avg. Session Duration	Bounce Rate	Bounces	Hits	Session Duration.1	Sessions	Unique Dimension Combinations
0	0	0,999230525505816	44152	44641	0	44186	44186
1	1,157407407407 41E-05	0	0	20	0,000115740740740741	10	10
10	0,000115740740740741	0	0	94	0,0048611111111111	42	42

Figura 98.

- **Session Duration:** durata della sessione
- **Avg. Session Duration:** tempo medio di durata delle sessioni
- **Bounce Rate:** la frequenza di rimbalzo è il rapporto tra le sessioni di una sola pagina divise per tutte le sessioni o la percentuale di tutte le sessioni nelle quali gli utenti hanno visualizzato solo una pagina e hanno attivato una sola richiesta al server Analytics.
- **Bounces:** un rimbalzo è una sessione di una sola pagina sul sito. In Analytics; un rimbalzo è calcolato in modo specifico come una sessione che attiva una sola richiesta al server Analytics, ad esempio quando un utente apre una singola pagina sul sito ed esce senza attivare altre richieste al server Analytics durante la sessione.
- **Hits:** un'interazione che si traduce in dati inviati ad Analytics. Ogni volta che il codice di monitoraggio viene attivato dal comportamento di un utente (ad esempio, l'utente carica una pagina su un sito web o una schermata in un'app mobile), Analytics registra tale attività. Ogni interazione viene inserita in un hit e inviata ai server di Google. Esempi di tipi di risultati includono:
  - hit di monitoraggio della pagina
  - hit di monitoraggio degli eventi
  - hit di monitoraggio e-commerce
- **Sessions:** numero di sessioni
- **Unique Dimensions Combinantions:** conta il numero di combinazioni valore-dimensione univoche per ciascuna dimensione in un rapporto. Ciò consente di creare la post-elaborazione delle dimensioni combinate (concatenate), che consente rapporti più flessibili senza dover aggiornare l'implementazione del monitoraggio o utilizzare slot di dimensioni personalizzate aggiuntive.

- **Social interaction:** raccolta delle interazioni degli utenti con i pulsanti social presenti all'interno del sito web.

Social Action	Social Entity	Social Network	Social Network and Action (Hit)	Social Type	Actions Per Social Session	Social Actions	Unique Social Actions
Share	https://www.certimetergroup.com/lavora-con-noi/svilupp	ShareThis_facebook	ShareThis_facebook : Share	Socially Engaged	1	1	
Share	https://www.certimetergroup.com/lavora-con-noi/svilupp	ShareThis_facebook	ShareThis_facebook : Share	Socially Engaged	1	1	
Share	https://www.certimetergroup.com/lavora-con-noi/svilupp	ShareThis_linkedin	ShareThis_linkedin : Share	Socially Engaged	1	1	

Figura 99.

  - **Social Action:** relativa azione intrapresa
  - **Social Entity:** entità social di riferimento
  - **Social Network:** si riferisce al Social Network adoperato
  - **Social Type:** tipologia di social
  - **Action Per Social Session:** numero di azioni avvenute per sessione
  - **Social Actions:** numero di azioni totali
  - **Unique Social Actions:** numero di azioni per utente

3.3 Data editor

Una volta presa conoscenza delle caratteristiche del data set, al fine che ne venissero identificate le peculiarità e le correlazioni principali, si è passati alla fase relativa alle trasformazioni dei dati a disposizione, ovvero i processi ETL.

Innanzitutto, è stata rivista la connessione alle fonti di origine, Google Analytics, mediante la verifica delle credenziali e la conferma della stabilità stessa; come spiegato nel paragrafo precedente, infatti, è stata anticipata la fase attinente alla creazione della reportistica ma con il solo intento conoscitivo del set di dati che, implicitamente, ha comportato il loro caricamento su Power BI Desktop. La convalida dell'effettiva salda e corretta connessione, è stata possibile mediante l'apposito comando "impostazioni origine dati" che ha ricondotto alla seguente schermata.

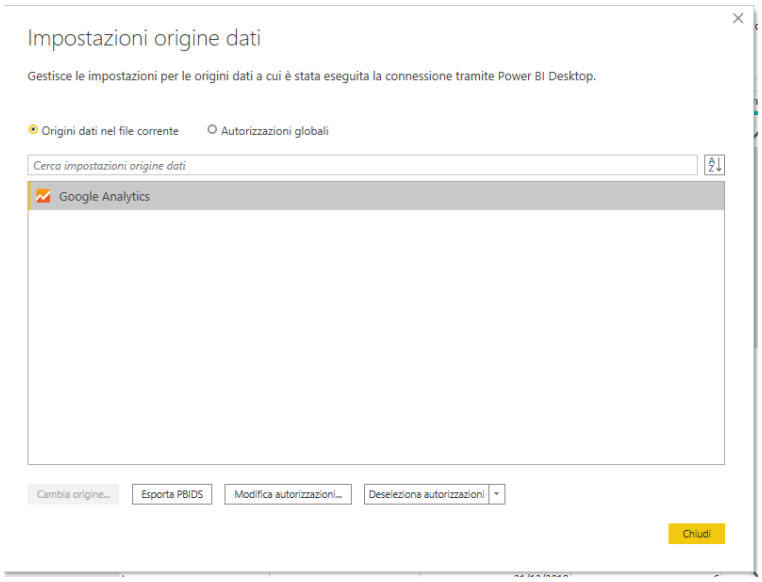


Figura 100.

Accertate le condizioni sul collegamento all'origine, ci si è mossi sull'effettiva manipolazione dei dati. Sono stati per primo inizializzate i gruppi dentro i quali aggregare le diverse tabelle. Per le originarie, appena estrapolate, è stata adoperata il gruppo di default denominato "Altre query".

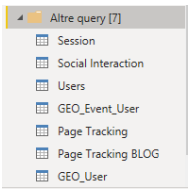


Figura 101.

Da questo momento, la focalizzazione è stata indirizzata alla modifica dei nomi di alcune colonne ed al loro ordinamento, oltre che alla corretta loro conversione nell'adeguato formato, conforme con i dati al proprio interno.

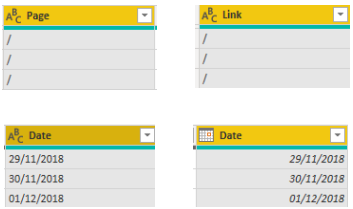


Figura 102.

Intento primario a questo punto, è stato l'assestamento della disposizione di specifiche informazioni, per l'agevolazione del filtraggio nei report finali. Nel dettaglio, è stato fatto

sì che l'URL delle singole pagine (es. /lavora-con-noi/sviluppatore-java), venissero estrapolate e suddivise a seconda del livello di nidificazione, identificato dal delimitatore "/" all'interno dell'indirizzo e, successivamente, assegnare ad ognuno dei livelli una dedicata colonna.

Ad esempio, dall'URL "/servizi/sviluppo-web" è stato possibile identificare la pagina "Servizi" (primo livello), a segnalare l'opportunità di risalita a ulteriori pagine quali proprio "Sviluppo Web" (secondo livello). Pertanto, è stato fatto in modo che la pagina "Servizi" venisse inserita all'interno della colonna, creata ad hoc, "Pagina 1" (ad indicare il primo livello, appunto) mentre la pagina "Sviluppo Web" nella conseguente colonna "Pagine 2" (ad indicare questa volta il secondo livello). Scopo di questa operazione è stata quella di agevolare, e quindi ottenere un incremento dell'efficienza, la ricerca delle singole pagine all'interno della reportistica. Il disporre di interi e completi indirizzi, non distinti per livello di nidificazione, ne avrebbe complicato l'analisi rendendola relativamente confusionaria, quanto non chiara e immediata.

Per la messa in pratica di questa trasformazione, sono stati in primis realizzati diversi gruppi di supporto rispettivamente chiamati: "Gerarchia pagine", "Duplicati" e "Ausilio", quest'ultimi nidificati all'interno del primo.

Il gruppo "Gerarchia pagine" prevedeva di raccogliere una tabella per ognuna delle pagine appartenenti al primo livello ("Home", "Chi siamo", "Contatti", "Servizi" e "Lavora con noi") con l'aggiunta della tabella "Ricerca" dedicata agli indirizzi scaturiti dalla digitazione nell'apposito form, proprio per ricercarne il contenuto all'interno del sito; in ultimo, ne è stata introdotta un'ulteriore con il nome di "Unione" per compattare le singole tabelle create in una unica tabella. Il gruppo "Duplicati" era invece destinato, come desumibile dal nome, ai duplicati delle tabelle originarie, in modo tale da rendere indipendente le trasformazioni in corso dai dati di riferimento. Il terzo e ultimo gruppo, "Ausilio", era finalizzato a tabelle di utilità momentanea.

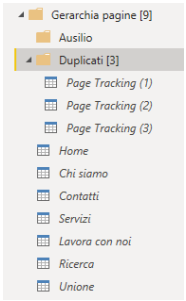


Figura 103.

Le operazioni sono avvenute sulla base della tabella “Page tracking” nel gruppo “Altre query”. Dopo le prime modifiche infatti, realizzate sul nome delle colonne, sul loro formato e sulla loro disposizione, si è cercato di distinguere i diversi livelli degli indirizzi, adesso presenti all’interno della colonna denominata “Link”. La funziona adoperata è stata la “Estrai” che ne ha permesso l’estrapolazione e lo spostamento, in una nuova colonna, della parte di testo voluto, in corrispondenza di uno specifico delimitatore.

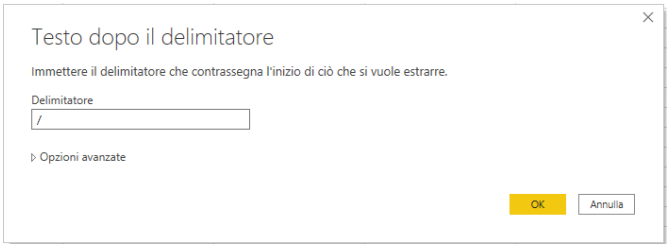


Figura 104.

In seguito a questo step, è stata rinominata questa nuova colonna in “Pagina 1” (ad indicare come detto il primo livello), oltre che essere stata riallocata, per sola chiarezza, in vicinanza della colonna “Link” stessa. Ultimo passaggio è consistito nella sostituzione delle righe vuote della nuova colonna, con il termine “Home” considerato che questa pagina non dispone di nidificazioni ed è individuabile dal solo e univoco indirizzo “/ “. . Analogo procedimento di estrapolazione, successivamente seguito dalle medesime modifiche (salvo la sostituzione delle righe vuote in “Home”), è stato attuato sulla colonna “Pagina 1” che ha comportato la creazione di una nuova colonna, denominata per semplicità “Pagina 2”, come il suo livello di nidificazione.

Link	Pagina 1	Pagina 2	Date	Unique Pageviews	Uniq
/lavora-con-noi	lavora-con-noi		24/01/2021	1	
/lavora-con-noi	lavora-con-noi		25/01/2021	2	
/lavora-con-noi	lavora-con-noi		26/01/2021	1	
/lavora-con-noi/sviluppatore-...	lavora-con-noi/sviluppatore-...	sviluppatore-android	07/12/2018	1	

Figura 105.

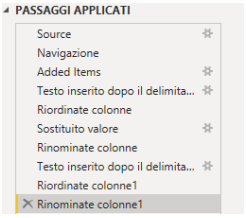


Figura 106.

Seppur completata la trasformazione, è stato necessario correggere alcuni aspetti. Le righe vuote nella colonna “Pagina 2”, a differenza di quanto accaduto per la colonna “Pagina 1”, non sono state sostituite con un valore a questo diverso. Ciò in quanto risultava impossibile sostituire l’insieme delle righe vuote con un preciso valore, una singola pagina pertanto. Se per gli spazi vuoti della colonna “Pagina 1”, infatti, si era consapevoli che provenivano dai valori “\ “ in “Link” che identificava a loro volta la Home, per gli spazi vuoti della colonna “Pagina 2”, invece, non si disponeva di tale certezza; le righe vuote, di fatto, potevano provenire da più dati quali ad esempio “contatti” o “chi-siamo”, privi di nidificazioni, a differenza invece dei dati come “lavora-con-noi/sviluppatore-android”. Era necessario sostituire ogni singola riga vuota con la rispettiva pagine di provenienza ma ciò non era consentito dalla funzione precedentemente adoperata di “Sostituisci valori”.

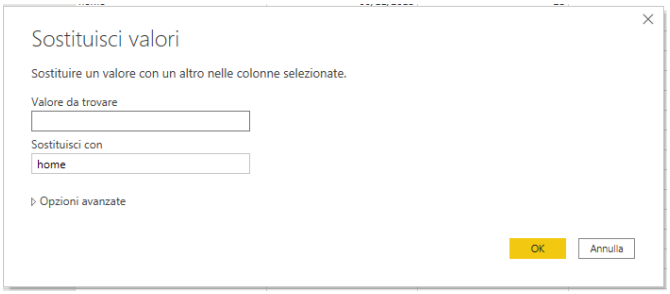


Figura 107.

Per risolvere tale problematica, è stato pensato di creare tabelle apposite per ogni pagina del primo livello. Questo è stato realizzato al fine di poter adoperare, a quel punto, la funzione “Sostituisci valori” con la consapevolezza di conoscere la pagina di provenienza delle righe vuote. Di non poca rilevanza, è stata però la necessità di inserire una colonna “Indici” nella tabella “Page tracking”, così da poter identificare

univocamente ogni singola riga al suo interno (e agevolare il successivo merge). La creazione delle tabelle destinate ad ogni pagina del primo livello è avvenuta per diverse esigenze e in due modi differenti.

Il primo si è basato sulla drastica copiatura dei dati filtrati dalla pagina “Page tracking”: nel particolare delle colonne “Indice”, “Pagina 1” e “Pagina 2”. Una volta creata e popolata, sono state sostituite le corrispettive righe vuote nella colonna “Pagina 2” e modificati piccoli aspetti, più che altro estetici.

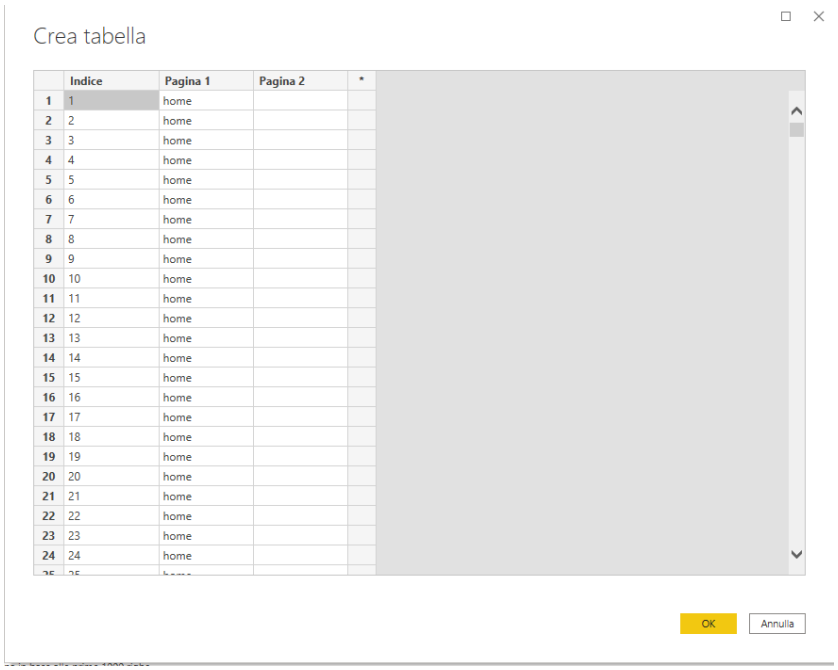


Figura 108.

Indice	Pagina 1	Pagina 2
1	Home	Home
2	Home	Home
3	Home	Home
4	Home	Home
5	Home	Home
6	Home	Home
7	Home	Home
8	Home	Home
9	Home	Home
10	Home	Home

Figura 109.

Il secondo metodo è avvenuto per esigenze computazionali; il numero elevato di righe non era tale da poter essere supportato dalla semplice operazione di copiatura e, pertanto, si è ricorsi all’operazione di “merge”. Il merging, altro non consiste che nel combinare due tabelle, ad esempio la “Page tracking” e la “Servizi”, quest’ultima costituita da una sola riga con indice non corrispondente alle righe effettivamente

attinenti agli indirizzi relativi a “Servizi” in “Page tracking”. Tutto ciò, al solo fine di poter convogliare le righe di “Page tracking” (filtrate sulle specifiche pagine richieste, in questo caso di “Servizi” appunto) nella tabella “Servizi”, mediante l’opzione “Right Anti”. Ovviamente, questo processo si è dovuto applicare anche per altre tabelle del gruppo “Gerarchia pagine” e, poiché l’operazione di “merge” dipende dal filtro in quel momento applicato alle pagine in questione (in questo caso “Page tracking” perchè base per le nostre trasformazioni), è stato essenziale realizzare delle duplicazioni della tabella “Page tracking” appunto, ognuna delle quali filtrate e dedicate a singole tabelle del gruppo “Gerarchie pagine”. Come per il primo metodo, sono stati in seguito introdotte modifiche organizzative ed estetiche.

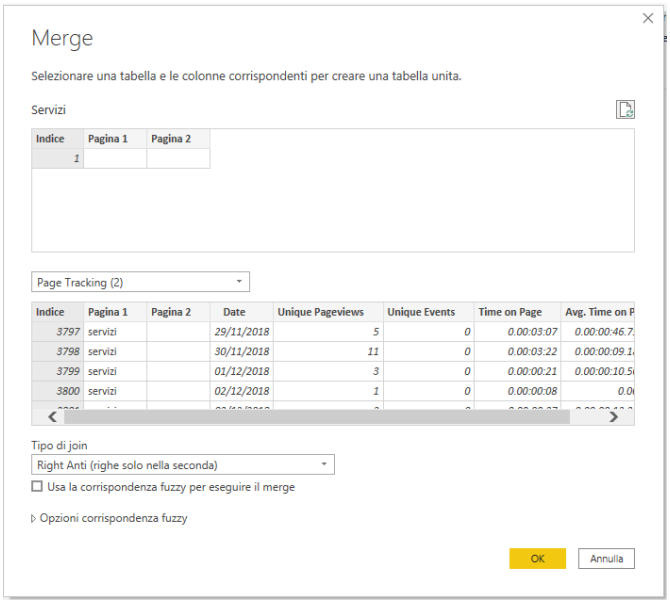


Figura 110.

Come ultima rilevante fase, sono state aggregate in un’unica tabella denominata “Unione”, le tabelle del gruppo “Gerarchia pagina”, mediante la funzione “Accoda”.

Figura 111.

In conclusione, si è realizzato il merge tra la tabella “Unione”, appena creata, e la “Page tracking” sulla base della colonna “Indice”, realizzando la tabella che verrà in seguito riportata nel model, ovvero la “Page tracking (M)”. Successivamente, si sono realizzati gli ultimi accorgimenti più sul piano visivo che altro.

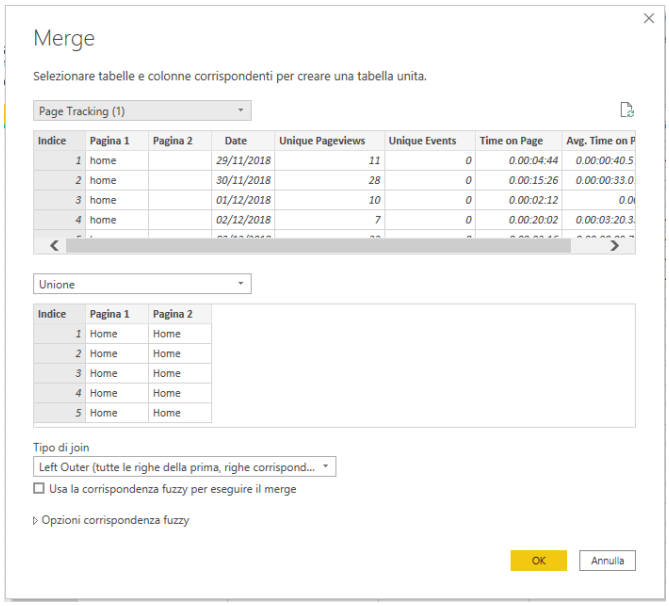


Figura 112.

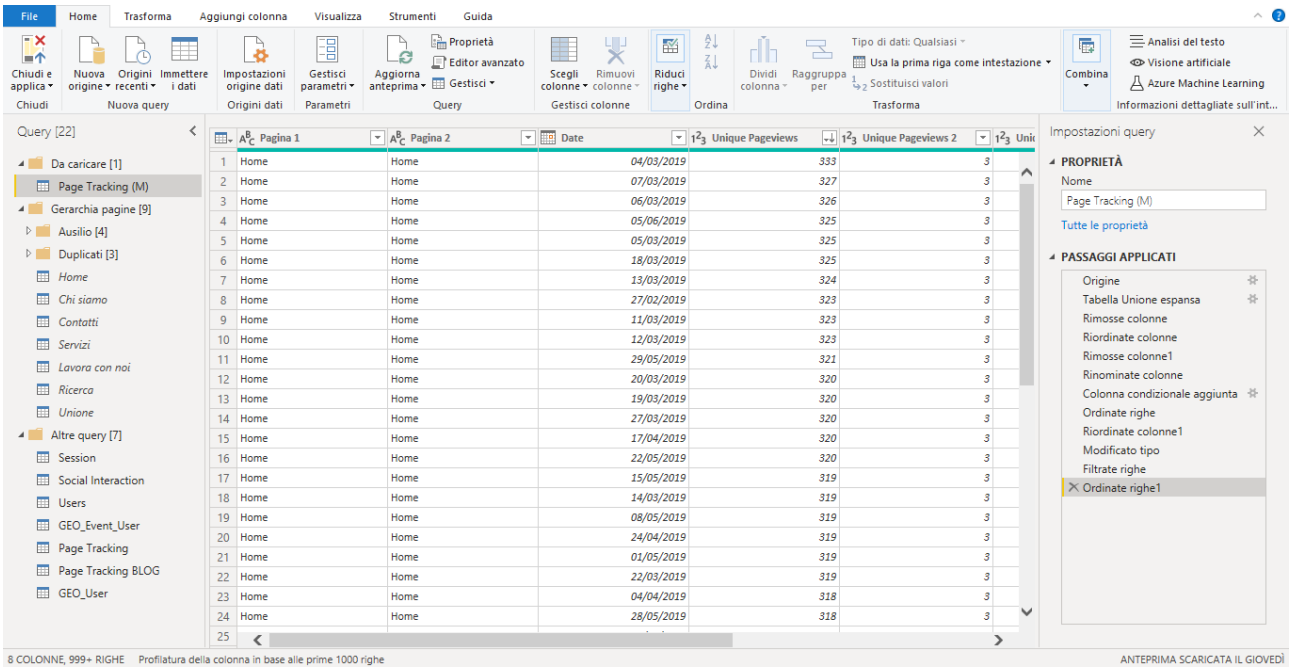


Figura 113.

3.2.1 Problematiche

Nei processi della business intelligence, come già spiegato nei capitoli precedenti, si riscontrano spesso diverse incoerenze o lacune all'interno dei data set; ciò conduce, di conseguenza, all'impossibilità di poter correttamente visualizzare le informazioni formalizzate, in quanto la distribuzione risulta non essere omogenea o del tutto corretta. Per rimediare, o semplicemente ridurre tali problematiche, si ricorre a diverse soluzioni. In linea di massima, solitamente, si cerca comunque di evitare la rimozione dei dati se non quando questi trasportano delle informazioni palesemente errate o incomprensibili. Non è però infrequente che, seppur vi siano evidenti incongruenze o anomalie fra due o più visualizzazioni determinate dall'utilizzo di una specifica gamma di dati, questa venga ugualmente adoperata, ma limitata e vincolata ad un determinato fine, interesse o intento dimostrativo.

Anche nel caso della società Alfa, sono emerse delle divergenze tra le illustrazioni prodotte dall'utilizzo dei medesimi dati ma, come spiegato, in seguito instradate e indirizzate in ottica e finalità diversa. L'esempio cardine è dato dalla colonna "Unique pageviews", nella tabella "Page tracking". Questa colonna sta ad indicare l'aggregazione delle visualizzazioni della pagina, generate dallo stesso utente durante la stessa sessione.

Una visualizzazione di pagina unica (unique pageviews per l'appunto) rappresenta il numero di sessioni durante le quali quella pagina è stata visualizzata una o più volte. Alcuni dei valori dell'anno 2019, e in particolare nel periodo compreso tra Marzo e Giugno, risultavano estremamente più elevati rispetto agli altri periodi in considerazione: un divario non trascurabile. Se mediamente si contavano da 1 a 10 visualizzazioni giornaliere, nel periodo in questione si toccavano valori che oscillavano dai 100 ai 300. Le interpretazioni a tale anomalia potevano essere ricondotte a due principali ipotesi: la prima, riguardante un incremento delle visualizzazioni dovuto agli effetti di una efficace campagna pubblicitaria; la seconda, invece, attinente ad un più plausibile numero di test di carico effettuati nel periodo in esame. Accertata l'infondatezza della prima ipotesi, è stata confermata la seconda. La soluzione scelta per la risoluzione del problema non è stata l'eliminazione dei valori anomali, bensì quella di conformarli al resto dei dati, considerata la conoscenza media del numero reale delle visualizzazioni. Questo è stato possibile mediante la funzione di "Colonna condizionale" che ha permesso la sostituzione degli specifici valori per mezzo di condizioni nidificate; ad esempio i valori superiori a 200 ma inferiori a 300, sono stati sostituiti con valori pari a 2. Ovviamenta sarebbe stato probabilmente più corretto inserire direttamente la media giornaliera, scelta che è stata scartata al solo fine dimostrativo di ulteriori funzioni presenti in Power BI. La nuova colonna è stata ridenominata "Unique pageviews 2".

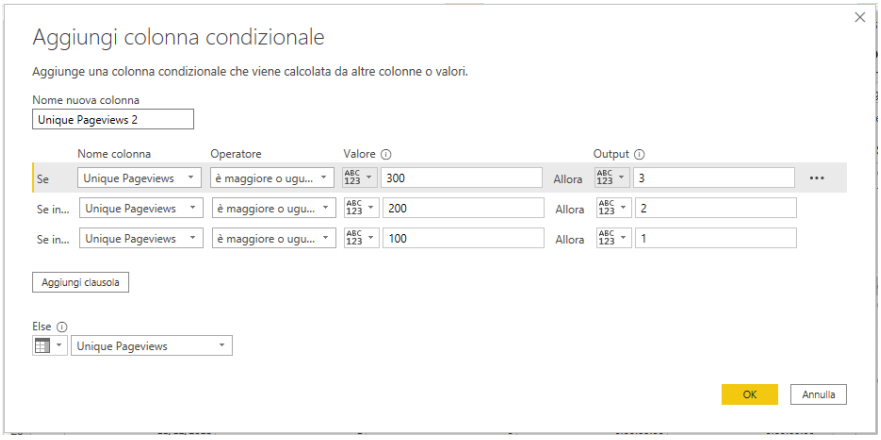


Figura 114.

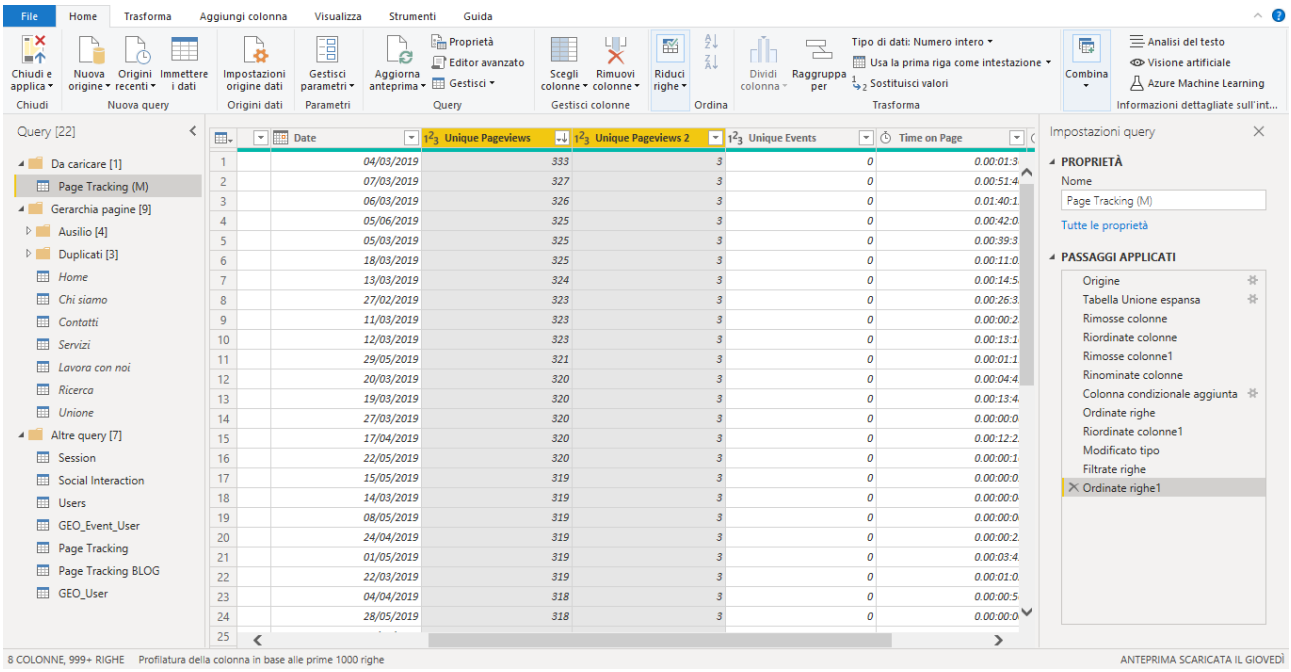


Figura 115.

L'esempio cardine è riscontrabile dal confronto tra due specifici report: da un lato, attinente al numero di visualizzazioni totali annue per tutte le pagine costituenti il sito web e, dall'altro, filtrate invece per la correlazione alla sola pagina denominata "Servizi". Come mostrato in figura, si nota che nel report riguardante le visualizzazioni annue sul complessivo set delle stesse pagine, vi sia una non trascurabile differenza tra i diversi anni. Il 2019 è infatti l'anno con il maggior numero di visualizzazioni, così sproporzionato che preseagisce che vi siano dei vuoti informativi nel resto delle altre annate. Nella figura B invece, attinente al secondo report focalizzato sulle pagine riguardanti gli specifi servizi offerti, si evince invece come, incoerentemente, le visualizzazioni dell'annata del 2020 siano in realtà superiori a quelle del 2019. Nonostante pertanto il primo report ci indicasse uno incolmabile divario tra i dati a disposizione nei diversi anni, il secondo mostra al contrario una caduta dell'ipotesi stessa.

3.4 Data model

Completata la fase di pulizia e organizzazione del data set, realizzata mediante il data editor di Power BI, si è passati alla sezione del data model: fase antecedente all'ultima

di reportistica. Lo scopo prefissato per questa fase, analogamente alla precedente appena vista, è l'incremento dell'efficienza e dell'esperienza dell'utente finale durante la vista delle visualizzazioni. Questo intento è messo in pratica non tramite il riordino dei dati, l'organizzazione e le relative trasformazioni, bensì mediante la creazione delle cosiddette "misure" o "nuove colonne"; questi calcoli, specificamente, aggiungono informazioni o ne estrapolano parte, ovviamente, dal data set a disposizione. Le misure, per mezzo del linguaggio DAX, consentono di delineare delle vere e proprie funzioni capaci di adeguarsi alle diverse condizioni imposte per mezzo dei report e, pertanto, costantemente mutabili a seconda delle esigenze. Le "nuove colonne", invece, non variano al mutare delle condizioni in quanto, seppur create e aggiunte al data set, rimangono a quest'ultimo vincolate per ogni sua relativa riga arricchendone, di conseguenza, le conoscenze [129].

Nel caso dell'azienda Alfa e delle relazioni tra i diversi aggregati di dati a disposizione, si è notata l'esigenza di delineare alcune ulteriori informazioni che descrivessero, con un maggior dettaglio, alcune considerazioni. Alcune di queste, hanno riguardato il fattore tempo e, nel particolare, attinente alla permanenza degli utenti sulle singole pagine e di seguito sull'intero sito. A tale proposito sono state realizzare diverse misure in supporto, tutte riguardanti la tabella "Page tracking (M)":

- **Ore complessive:** riporta il numero totale di ore trascorse su specifiche o più pagine. Non è stata utilizzata la funzione a disposizione "Time" in quanto il quantitativo di ore complessive, nel caso di selezione multipla di mesi o annate, avrebbe superato il limite delle 24 ore e non sarebbe quindi stato possibile poter correttamente visualizzare il risultato nel formato "hh:mm:ss". Nella pratica, sono stati prima convertiti i minuti ed i secondi in ore e, successivamente, si è fatta la somma di questi.

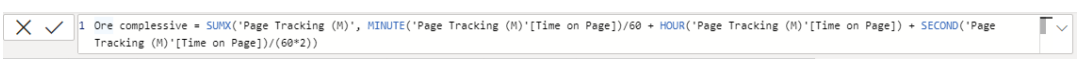


Figura 116.

- **Tempo totale:** restituisce il tempo totale ma in secondi, al fine di disporre di un ulteriore grado di precisione. È una misura principalmente indirizzata ad una successiva, ovvero il "tempo medio". Anche in questo caso, per il medesimo motivo visto per "ore complessive", non è stata adoperata la funzione "time" offerta dal linguaggio DAX.

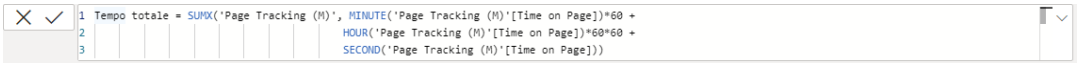


Figura 117.

- **Tempo medio:** fornisce il tempo trascorso mediamente sulle pagine e periodo di tempo selezionati dalla sezione reportistica. In questo caso è stata utilizzata la funzione "time" ed è stata sfruttata la misura "tempo totale".

Figura 118.

- **Tempo totale (time):** con lo stesso scopo della misura "tempo totale", gli differisce per l'utilizzo della funzione "time" e quindi non restituire un solo numero intero (i secondi) ma un formato espresso in ora, con limite però delle sole 24 ore.

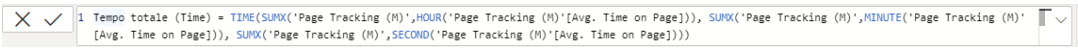


Figura 119.

Ultimo contesto dove risulta applicato i calcoli offerti nel data model, è stato quello relativo alla tabella "GEO_user" e, nel dettaglio, per la relazione tra visite e collocazione geografia degli accessi.

- **Percentuale visitatori:** l'opzione intrapresa per questa informazione non è stata la "misura" bensì la "nuova colonna": questa altro non indica che la percentuale del traffico relativo ad ogni luogo dato dal rapporto tra specifiche visite riguardanti le righe selezionate e la somma totale di visite.

Figura 120.

City	Country	Latitude	Longitude	Regione	Unique Pageviews	Percentuale visitatori
Turin	Italy	45.0703	7.6869	Piedmont	2000	57,50%
Milan	Italy	45.4642	9.1900	Lombardy	383	11,01%
Ravenna	Italy	44.4184	12.2035	Emilia-Romagna	97	2,79%

Figura 121.

3.5 Relazioni

Passaggio intermedio tra le sezioni riguardanti il data model e la reportistica, sono le relazioni tra le differenti tabelle. Nel caso della società Alfa, questa è stata pressocchè ininfluente in quanto i diversi aggregati di dati risultavano indipendenti l'uno dall'altro. Unica eccezione ha riguardato le sole coordinate geografiche tra le tabelle "GEO_User" e "GEO_Event_User", che rispettivamente riguardano il luogo di solo accesso ed il luogo di accesso con interazione.

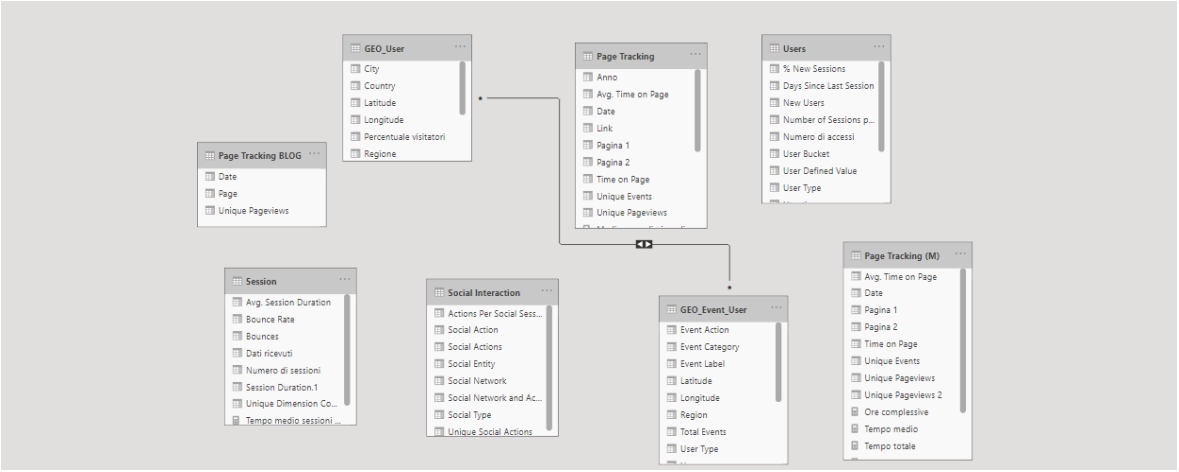


Figura 122.

3.6 Report

Terminate le fasi inerenti all'elaborazione e alla gestione dei dati, si è passati alla parte pertinente alla reportistica. La sua finalità è stata la rappresentazione per via grafica dei dati e delle loro relazioni, al fine di generarne una conoscenza accessibile e fruibile. Intento principale, del resto, era la realizzazione di uno strumento di supporto allo stadio decisionale della società Alfa, che potesse pertanto conferire un valore aggiunto alle scelte e all'interpretazione dei diversi contesti e ambiti, costituenti e coinvolti.

Con lo scopo di ottimizzare l'esperienza visiva e facilitarne la comprensione, sono state realizzate sei differenti pagine, ognuna delle quali inerenti ad una specifica tematica e con propri e dettagliati report. Le pagine e/o tematiche distinte, sono state:

- **Presentazione:** è la pagina che consente di conferire una visione generale della realtà Alfa. Da questa, infatti, è possibile rilevare il numero di visualizzazioni sull'intero sito per anno, le località nelle quali si sono verificati gli accessi e un confronto rapido delle pagine con la maggiore percentuale di traffico.

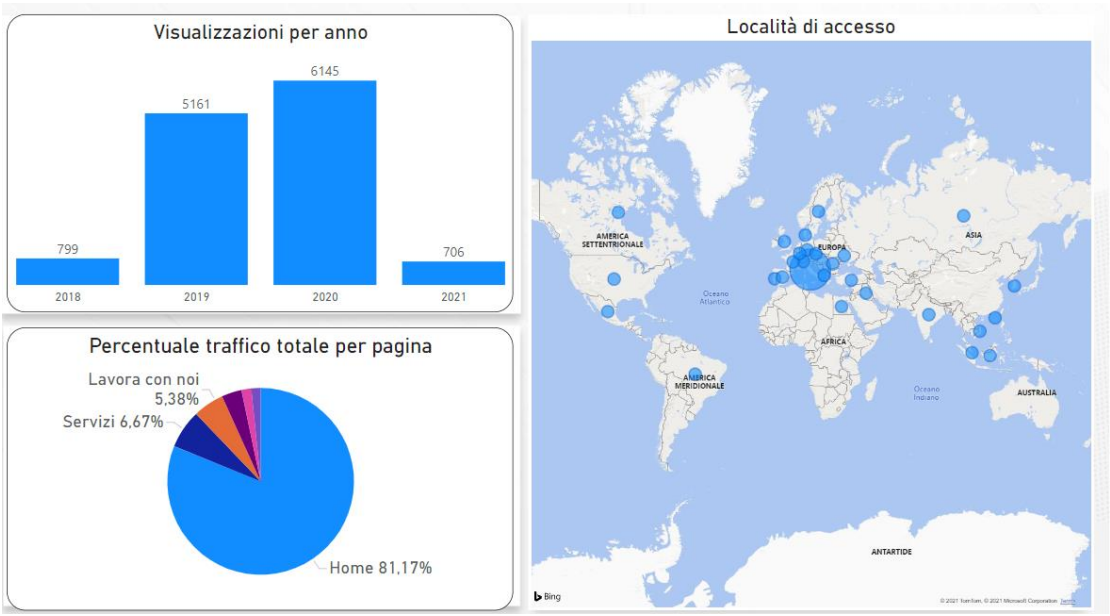


Figura 123.

- **Geografia:** in questa pagina si è cercato di descrivere le caratteristiche del mercato italiano, il più rilevante, mediante la focalizzazione sulla percentuale delle visualizzazioni per singola città e la raffigurazione delle dimensioni delle regioni con più interazioni avvenute all'interno del sito.

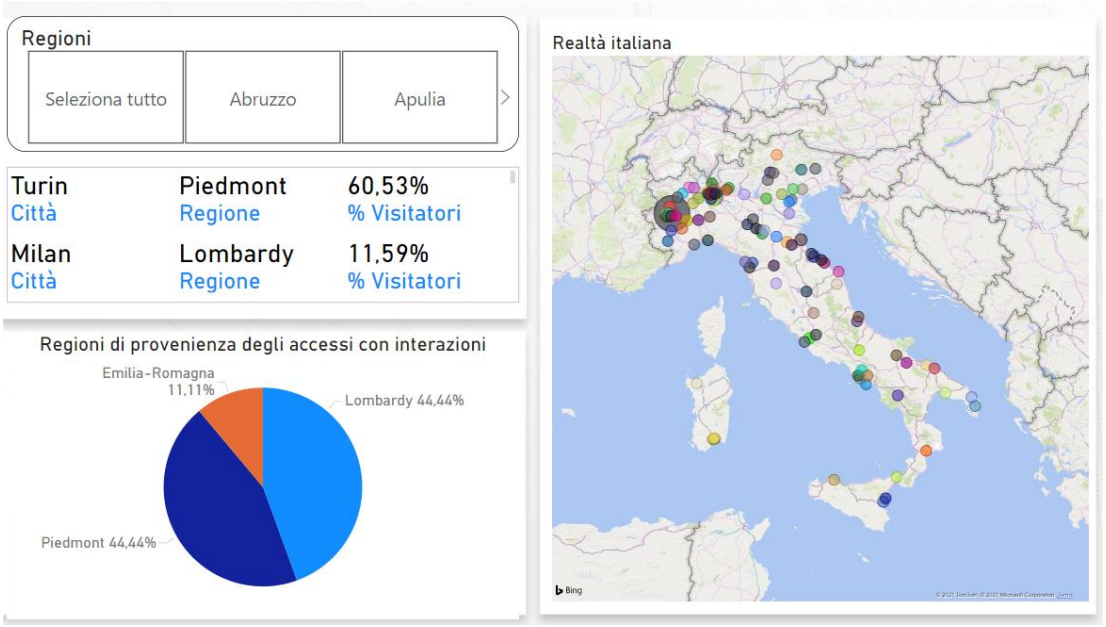


Figura 124.

- **Visualizzazioni:** la tematica qui affrontata è stata basata sul confronto delle visualizzazioni tra le singole pagine, di primo e secondo livello. Ciò è stato rappresentato sia in chiave percentuale che sui valori effettivi e numero medio. In aggiunta, è stato introdotto un grafico che rappresentasse l'andamento nei singoli mesi.

Figura 125.

- **Permanenza:** aspetto non poco sensibile relativo ad un sito web, è la comparazione tra la permanenza sulle singole pagine ed il numero stesso delle loro visualizzazioni. In questa pagina di reportistica è stato fatto sì che ciò potesse essere messo in pratica. Sono riportate le ore complessive, il tempo medio ed il confronto, sia in percentuale che in valori effettivi, tra il tempo speso sulla pagine e le loro visualizzazioni.

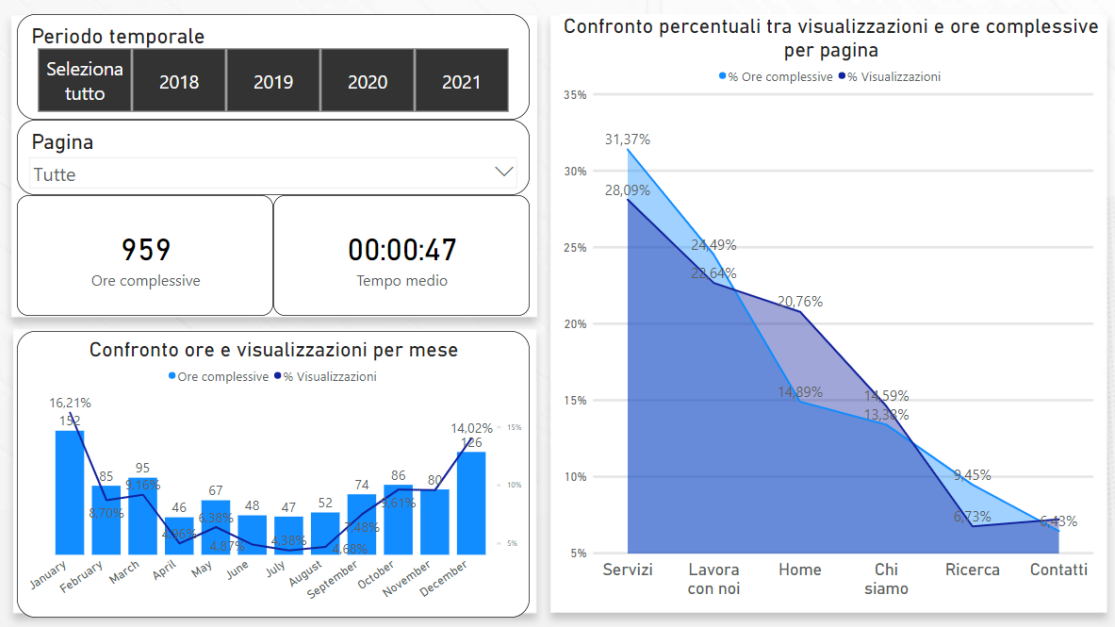


Figura 126.

- **Servizi:** aspetto cardine della società Alfa, sono i servizi offerti ai propri clienti. È stato pensato, pertanto, di proporre una pagina con diversi grafici, tali che riportassero i raffronti fra le visualizzazioni tra i singoli servizi offerti nelle diverse annate e la distribuzione del tempo trascorso all'interno delle loro pagine. Sono stati inseriti ulteriori informazioni sul tempo totale ed il tempo medio, per conferire un riferimento sulle distanze tra le misure.

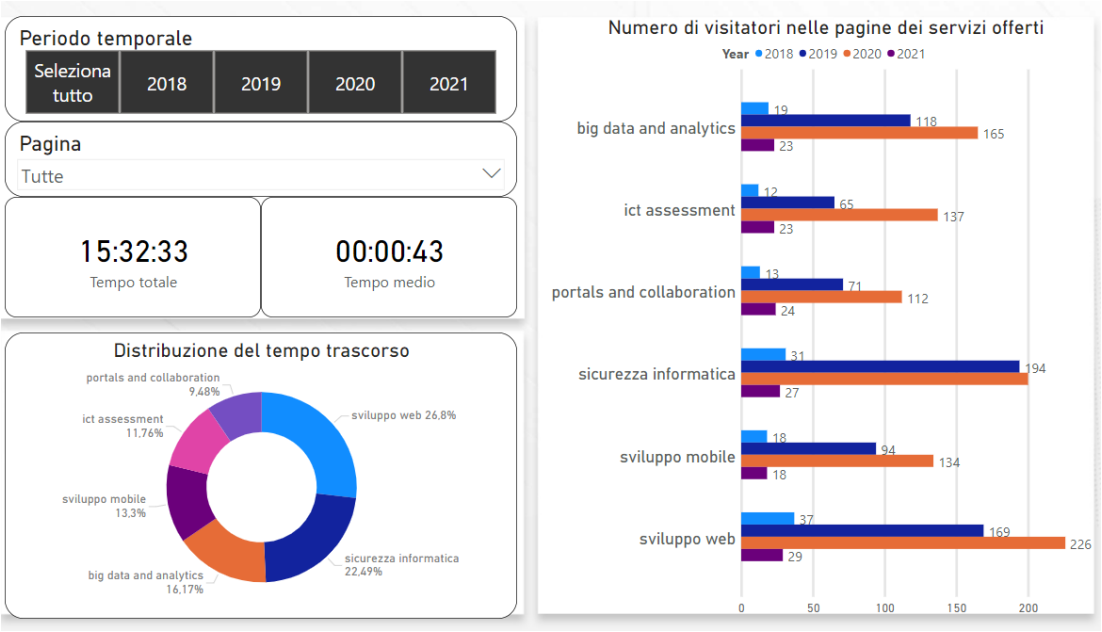


Figura 127.

- **Utenti:** un sito dispone di utenti che possono essere nuovi o già con almeno un avvenuto accesso. In questa pagina è stato ritratto il comportamento di coloro che hanno frequentato le pagine del sito, riproducendone il quantitativo di rientrati e nuovi, il loro numero per ogni range di permanenza ed il loro numero per tempo trascorso dall'ultimo accesso. In questo modo, è stato possibile tracciare un'idea dell'interesse da parte dei visualizzatori.

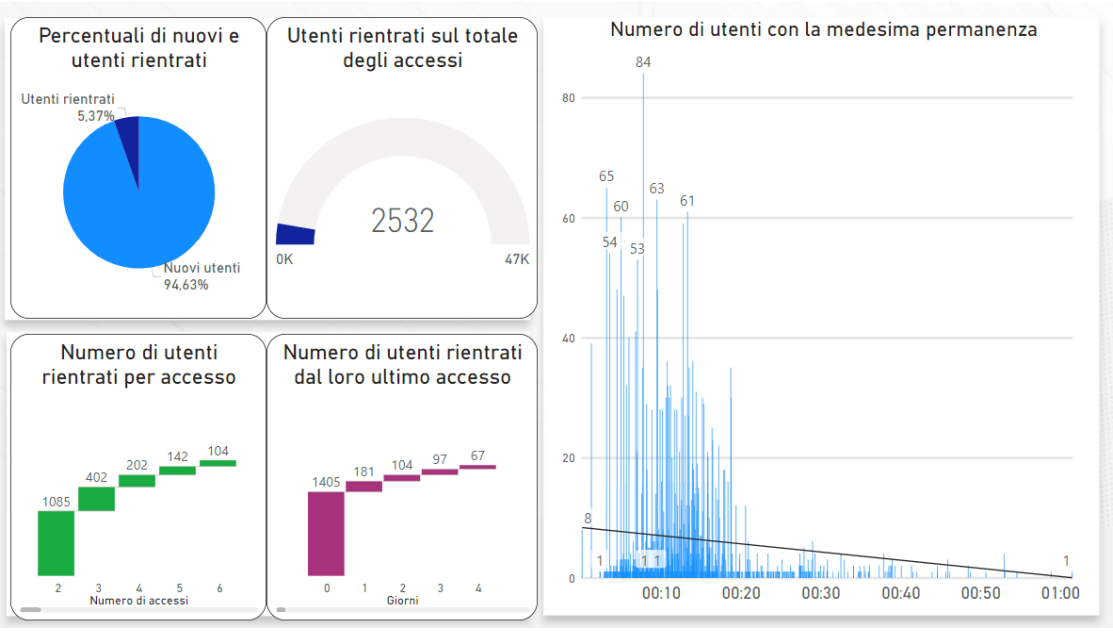


Figura 128.

3.7 Risultati e considerazioni

Una volta completate le differenti fasi riguardanti la preparazione, l'organizzazione e la proiezione dei dati verso una loro forma più "conforme" ad un'interpretazione prettamente più funzionale ai fini aziendali, si è passati alla successiva fase di analisi dei risultati emersi. In questo stadio conclusivo, destinato all'area decisionale della società, si fa sì che si possano elaborare delle informazioni più compatte e definite, tali da poter permettere di intraprendere o meno determinate strade, per una più adeguata direzione di precisi ambiti, strategie e/o operazioni. È mediante la reportistica che si ha modo di valutare lo stato di salute dell'azienda o società in questione e, pertanto, di giungere a delle conclusioni utili.

Nel caso della società Alfa, è stato possibile trarre già diverse considerazioni sin dalla pagina più generica. Non focalizzata su alcuna precisa tematica, infatti, la pagina cosiddetta di "presentazione" ha permesso di conoscere la realtà nella quale è immersa la società in questione. Si è constatato come il mercato di riferimento fosse situato anzitutto in Europa, in particolare in Italia, dato il numero non comparabile di accessi qui avvenuti rispetto al resto degli altri stati. Un modesto interesse è stato riscontrato anche in altri continenti, ma in dimensioni tali da non poter assumere connotazioni significative. È emersa invece una crescita significative nel 2020 rispetto all'anno precedente del traffico sull'intero sito, passata da un totale di visualizzazioni di 5161 a ben 6145, corrispondente ad una crescita pari a circa il 19%. Il 2018 ed il 2021, come già spiegato nel paragrafo "Problematiche", non possono essere ritenuti significativi per via del limitato numero di dati disponibili. Il 2018 dispone solo del periodo che va da Ottobre a Dicembre, mentre il 2021 del solo mese di Gennaio. Ulteriore chiara informazione attinta, è stata la comparazione in percentuale del numero di visualizzazioni per le pagine di primo livello, ovvero le pagine con possibili nidificazioni al proprio interno. In ordine decrescente si hanno avuto: "Servizi", "Lavora con noi", "Home", "Chi siamo", "Contatti" e "Ricerca". Il fatto che la pagina "Home" non sia stata la più cliccata è stato interpretato come mancata popolarità del marchio Alfa. Si è ritenuto, infatti, che gli utenti siano giunti

al sito tramite ricerche indirette e, conseguentemente, atterrando principalmente in pagine riguardanti i servizi offerti e le posizioni aperte. Non a caso, difatti, proprio queste due pagine disponevano di contenuti comuni a tante altre società del medesimo mercato.

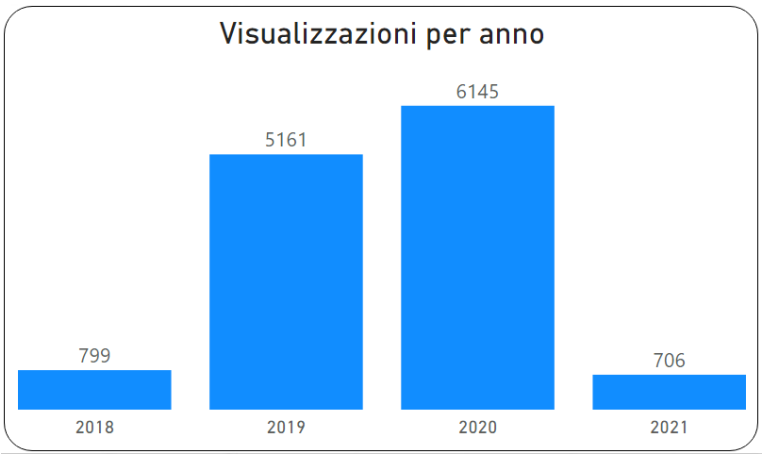


Figura 129.

Figura 130.

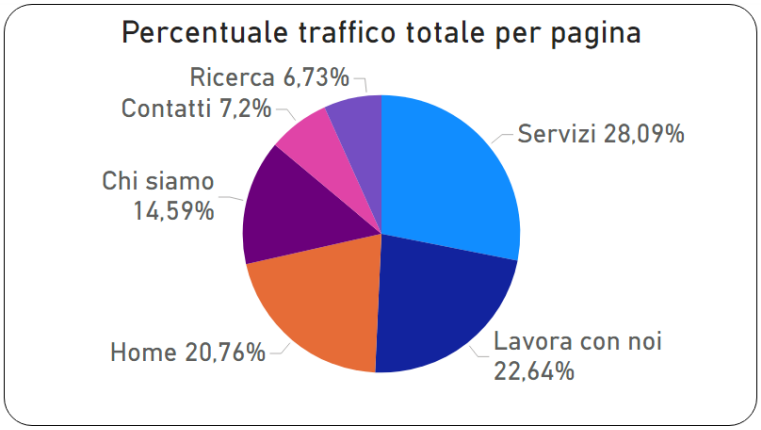


Figura 131.

Una volta preso atto che quello italiano è il mercato di riferimento della società Alfa, è stato deciso di discernere le peculiarità territoriali. Nel particolare, è stato constatato che la società, con sede a Torino, abbia anche interessi fuori dagli stretti confini comunali e della città metropolitana, estendendosi nell'intero territorio piemontese e limitrofo. Il comune di Torino ed il suo hinterland è, per ovvie ragioni, l'area che registra il maggior numero di accessi, coprendo circa il 60% del totale. Seguono Milano e Ravenna, rispettivamente in Lombardia ed Emilia-Romagna, entrambe regioni che fanno registrare – dopo il Piemonte - il numero più elevato di accessi con interazioni al sito, a comprova di una significativa attenzione verso Alfa da parte di potenziali clienti locali.

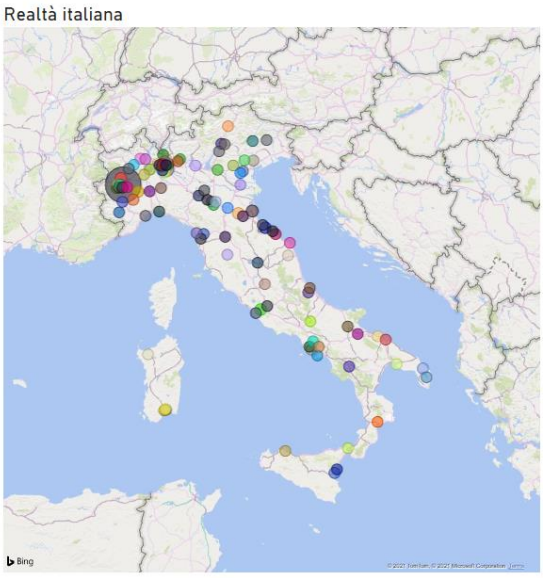


Figura 132.

Turin Città	Piedmont Regione	60,53% % Visitatori
Milan Città	Lombardy Regione	11,59% % Visitatori
Ravenna Città	Emilia-Rom... Regione	2,94% % Visitatori

Figura 133.

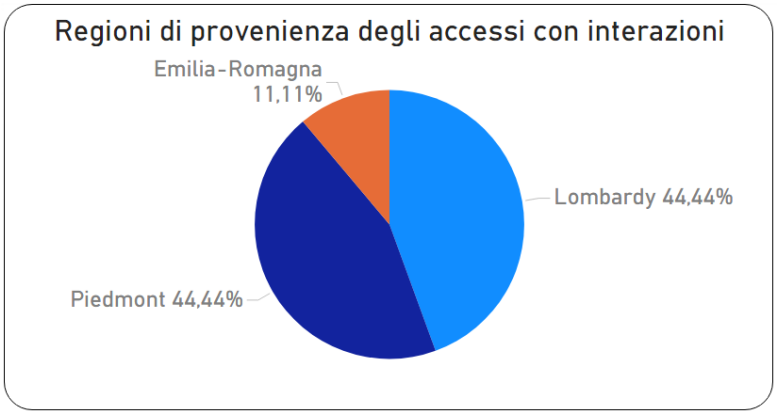


Figura 134.

Il numero delle visualizzazioni, come visto dai grafici prettamente più generici sulla realtà Alfa, sono stati diversi per periodo e anche a seconda della pagina in considerazione. Un successivo confronto, adoperando le variabili di tempo e pagina di riferimento, è stato realizzato sul terzo aggregato di immagini, al fine di delinearne andamenti e caratteristiche. Escludendo gli anni 2018 e 2021 per i motivi già citati, è stato possibile evidenziare una proporzione pressocché stabile della distribuzione delle visualizzazioni dal 2019 al 2020. Vi sono state comunque delle differenze, ma non significative. Le pagine “Servizi”, “Lavora con noi” e “Chi siamo” hanno visto una piccola crescita a discapito delle pagine “Contatti” e soprattutto “Home”. Questo ha potuto significare un aumento della visibilità del sito di Alfa, possibilmente anche legato ad una crescita della clientela per la quale l’azienda ha prestato servizio o di modifiche rilevanti all’interno del sito.

Pagina 1	% Visualizzazioni
Servizi	27,48%
Lavora con noi	22,11%
Home	21,20%
Chi siamo	14,47%
Contatti	8,22%
Ricerca	6,53%

Pagina 1	% Visualizzazioni
Servizi	28,49%
Lavora con noi	23,38%
Home	20,15%
Chi siamo	15,31%
Contatti	6,72%
Ricerca	5,94%

Figura 135.

Se le proporzioni sono rimaste le medesime tra i due anni, il numero effettivo totale di visualizzazioni avvenute è invece aumentato. Se nel 2019 sono state 5161, nel 2020 sono passate a 6145 ma mantenendo una media giornaliera di accessi analoga.

Figura 136.

Discorso differente può essere espresso per quanto concerne invece le visualizzazioni per singoli mesi o in piccoli aggregati. Per entrambe le annate si nota una simile riduzione del traffico nel secondo quadrimestre, corrispondente al periodo estivo. Del resto, questo risulta un arco di tempo in cui normalmente si verifica una diminuzione sia di erogazione che di richiesta di in svariati settori. Se nei mesi più caldi si è riscontrato un comportamento simile sul comportamento degli utenti, non si può sostenere la medesima constatazione rispettivamente per il primo e terzo quadrimestre. Nel 2019 infatti, a differenza che del 2020, il primo quadrimestre ha registrato il quantitativo di connessioni maggiore mentre, al contrario, il 2020 ha registrato il suo picco nel terzo quadrimestre. L'interpretazione a tale fenomeno è stato associato alla pandemia di covid-19 che, già a partire dai primi mesi del 2020, ha rallentato - se non interrotto - innumerevoli rapporti e messo in difficoltà molte medio-piccole realtà, clienti di riferimento di Alfa. D'altro canto, negli ultimi mesi del medesimo anno, abbiamo assistito ad una qual certa ripresa del normale svolgimento delle attività produttive che ha fatto da contraltare in buona parte alla crescita emersa. Si aggiunge che l'aumento in termini assoluti del numero delle visualizzazioni, può essere derivato anche da una maggiore

occupazione in remoto grazie allo smart working adottato dalla quasi totalità di aziende di servizi.

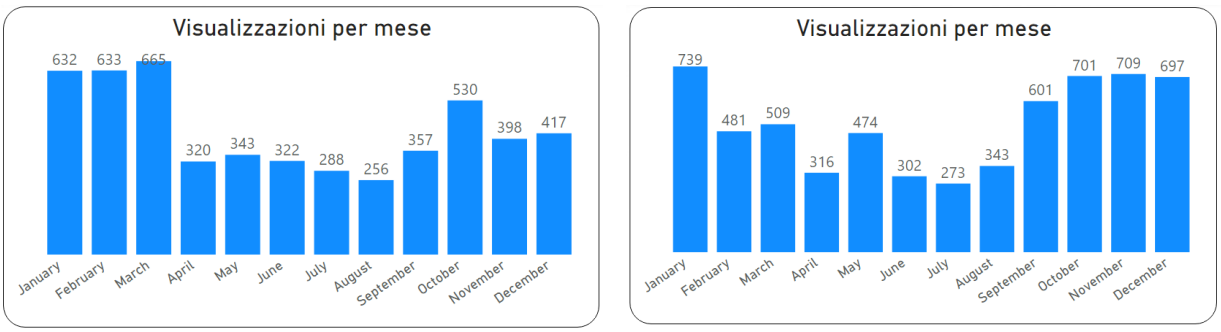


Figura 137.

Ulteriori confronti sono stati ovviamente possibili, nel particolare anche per quanto riguarda le pagine del secondo livello. Da questi grafici è stato infatti possibile confrontare eventuali variazioni dell'interessamento verso particolari posizioni e servizi stessi, e soprattutto in quale periodo dell'anno ciò è avvenuto.

Figura 138.

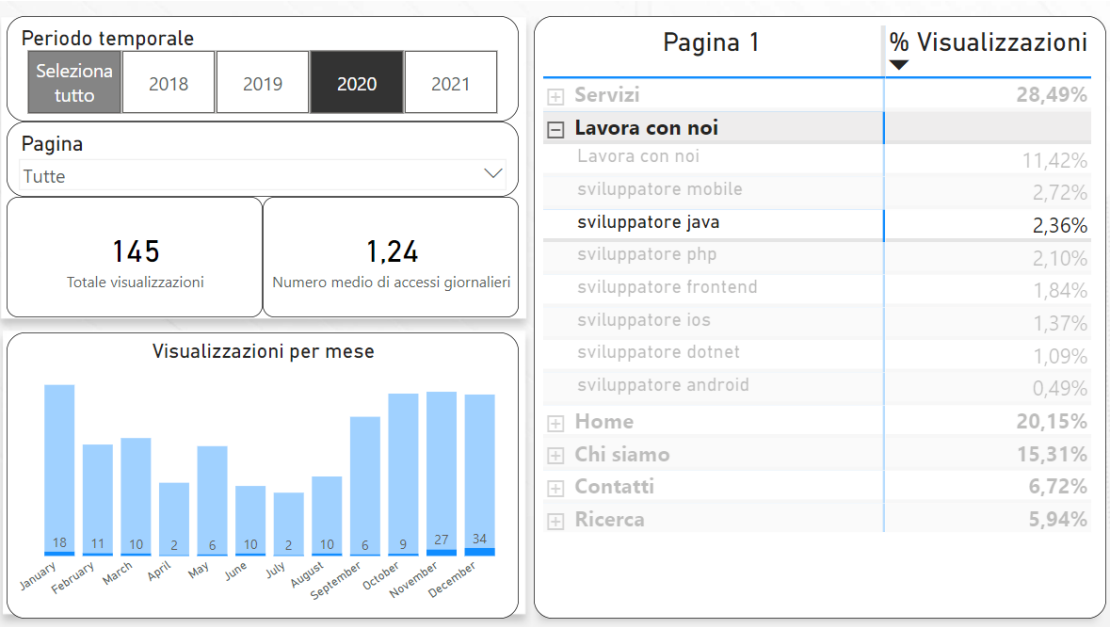


Figura 139.

Ulteriore aspetto che si è potuto rilevare, è stata la consistenza tra il data set e la reportistica, anche nelle lacune di dati come già accennato per gli anni 2018 e 2021.

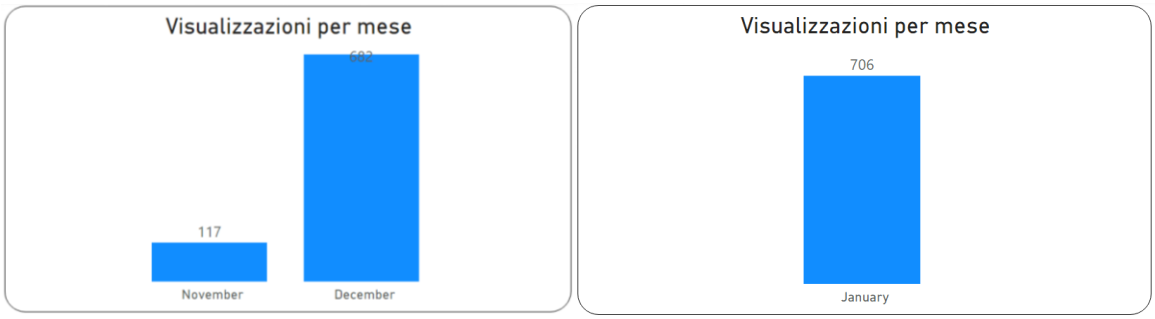


Figura 140.

Volendosi spingere oltre, è stata fatta un'osservazione sulla possibile previsione del traffico del 2021 di cui si dispone solo il mese di gennaio, confrontandolo con il rispettivo del 2020. Si nota una riduzione delle visualizzazioni, ma decisamente contenute se relazionata al difficile contesto pandemico ancora presente nel 2021.

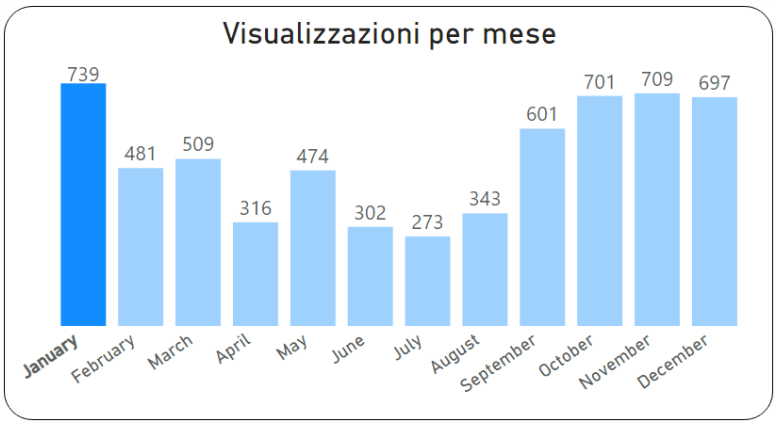


Figura 141.

Informazione non del tutto ininfluente, utile per una migliore valutazione delle prestazioni del sito, è stata la permanenza degli utenti sulle rispettive pagine e quindi il tempo da ognuno speso su queste. Tale dettaglio infatti, insieme alle visualizzazioni, ha consentito la delineazione di ulteriori aspetti comportamentali degli utenti che hanno effettuato l'accesso, interpretati come effetto e reazione provocati dalla struttura stessa del sito. Di conseguenza, per il realizzo di tale analisi, sono stati adoperati i report della pagina cosiddetta "Permanenza". Da questa, pertanto, è stato fatto sì che si potessero confrontare, in percentuali e valori effettivi, le distinte cifre per pagina rispettivamente del tempo trascorso e delle relative visualizzazioni avvenute. A differenza però di quanto visto fino adesso, non sono stati esclusi gli anni 2018 e 2021 (ma in questo caso per il solo confronto in percentuale) poiché l'analisi non ha richiesto una relazione al periodo temporale di riferimento dei dati. Le percentuali dei due anni sono state considerate attendibili e quindi incorporate nello specifico grafico. Quest'ultimo, successivamente, ha mostrato quanto alcune pagine mostrassero valori divergenti per le due variabili. Nel particolare, la pagina "Home" ha riportato una percentuale delle visualizzazioni pari al 20,63% del totale ma solo il

15,06% delle ore complessive, a significare quanto relativamente poco fosse d'interesse la pagina. Più comparabili invece, le pagine "Servizi", "Lavora con noi", "Chi siamo" e "Contatti" che hanno presentato sì dei discostamenti ma non significativi; le percentuali sulla permanenza sono risultate sempre al di sopra, a dimostrazione di una maggiore propensione degli utenti a dedicarvi maggiore tempo rispetto alla pagina "Home".

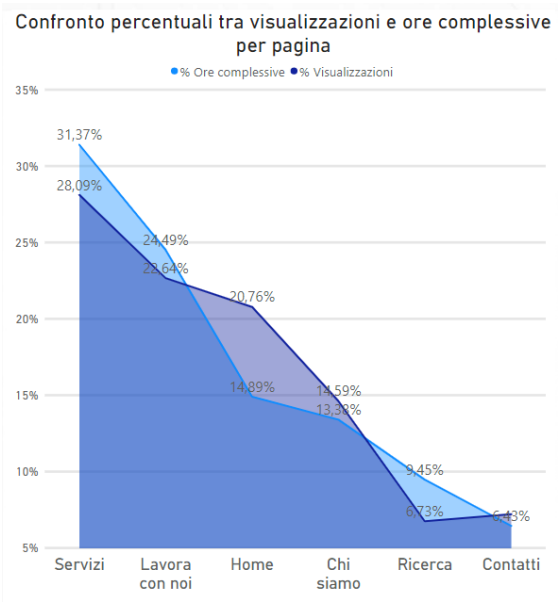


Figura 142.

Per il confronto tra le ore effettive e la percentuale delle visualizzazioni, è stato necessario non considerare gli anni 2018 e 2021 in quanto la comparazione è avvenuta sulla base delle mensilità. In questo modo, mantenendo in percentuale le visualizzazioni, è stato possibile poter confrontare più agevolmente le due misure e denotare l'andamento ed il comportamento degli utenti nei diversi momenti dell'anno. Ciò ha permesso di dedurre come il trend delle visualizzazioni fosse il medesimo delle ore spese, non facendo pertanto supporre eventuali situazioni anomale. Come già denotato dai grafici precedenti, i mesi estivi sono risultati i meno trafficati dell'anno.

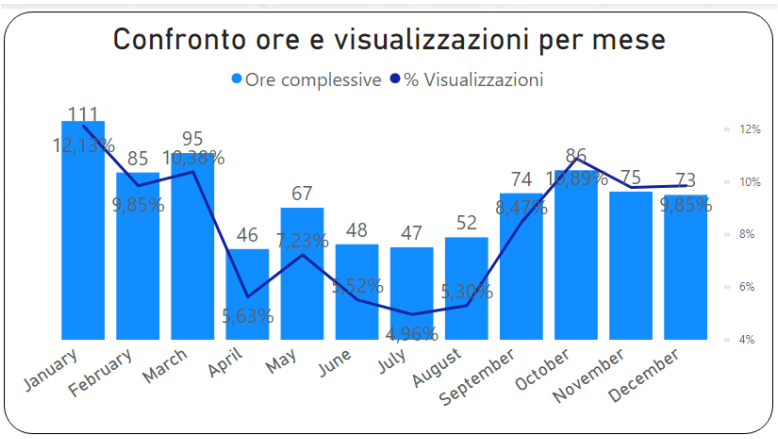


Figura 143.

La società Alfa, inserita nel mercato della consulenza informatica indirizzata alle piccole o medio/grandi realtà, si occupa di offrire servizi dedicati quali possono risultare lo

sviluppo web o la sicurezza informatica. L'affluenza sulla pagina dedita ai diversi servizi elargibili, pertanto, può e ha permesso di distinguere, o comunque disegnare, una traccia delle preferenze dei clienti. La reportistica chiamata "Servizi", per l'appunto, ha dato modo di rilevare le propensioni degli utenti, anche consentendo di porre o meno dubbi sull'efficienza della struttura stessa del sito. Sempre per via della carenza di dati nei diversi mesi degli anni 2018 e 2021, sono stati considerati i soli 2019 e 2020 al fine di comparare l'afflusso sui diversi servizi, nella pratica sulle pagine nidificate o in questo caso di secondo livello. Si è constatato che il 2020 ha ottenuto, anche come già conseguito dalle precedenti deduzioni, un incremento delle visualizzazioni per ogni servizio. Nel dettaglio, il servizio di "ict assessment" ha più che raddoppiato le visualizzazioni rispetto al 2019, sottolineando l'importanza che ha guadagnato la realizzazione dei percorsi di analisi per il perseguimento degli obiettivi dei clienti. Lo "sviluppo web" è stato il servizio con il maggiore afflusso del 2020 superando, rispetto al 2019, la "sicurezza informatica" che ha visto sì una crescita, ma non rilevante. Stabili nelle posizioni del 2019, e quindi con proporzionali crescite, rispettivamente i servizi di "big data e analytics" e di "sviluppo mobile", mentre è da sottolineare il dato relativo al "portals and collaboration" che ha evidenziato apprezzabili margini di crescita.

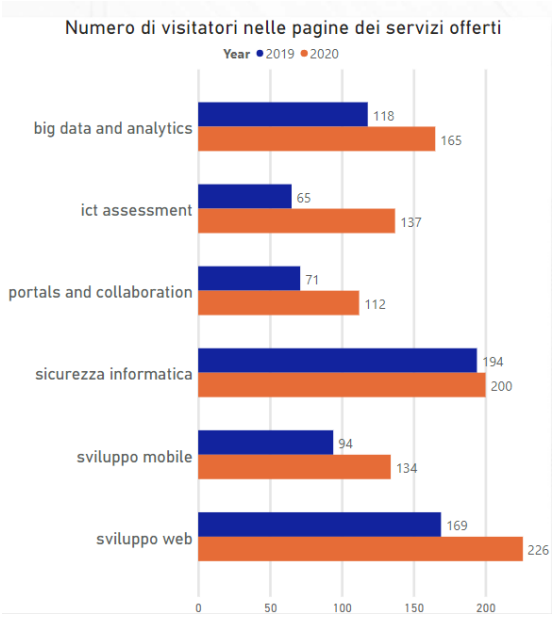


Figura 144.

Le percentuali del tempo trascorso confermano la rilevanza del servizio dello “sviluppo web” rispetto gli altri, con il 26,31% del tempo totale; non a caso risulta precedere quello attinente alla “sicurezza informatica”, a sua volta seguito dal servizio “big data and analytics” , e via discorrendo.

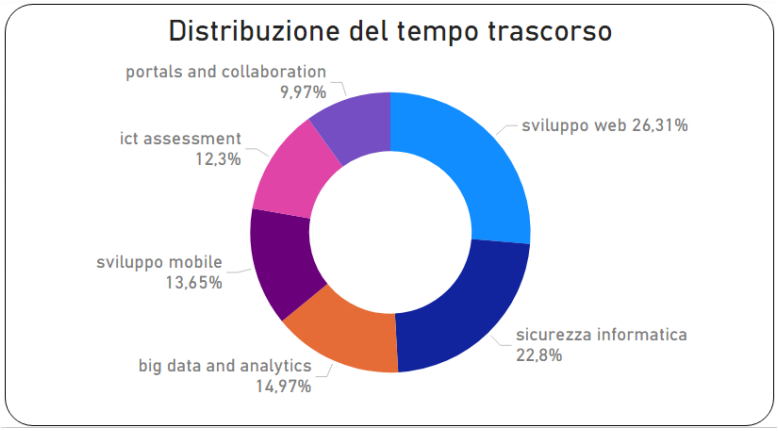


Figura 145.

L’ultimo set di grafici ha riguardato le tipologie degli utenti, basato principalmente sulla distinzione tra utenti rientrati e utenti al loro primo accesso. Questo confronto ha permesso di determinare una percentuale pari al 5,37% degli utenti rientrati rispetto al 94,63% di utenti al loro primo accesso. Ciò indica come sia di carattere prioritario valutare l’esigenza di procedere ad una rivisitazione del sito, con lo scopo di renderlo più attrattivo, tale che possa invogliare gli utenti ad un successivo rientro o, ancor più efficacemente, assicurare un soddisfacimento facilmente perseguibile del potenziale cliente sin dal suo primo ingresso. In valori effettivi, gli utenti rientrati corrispondono a 2532 individui su un totale approssimabile di 47000.

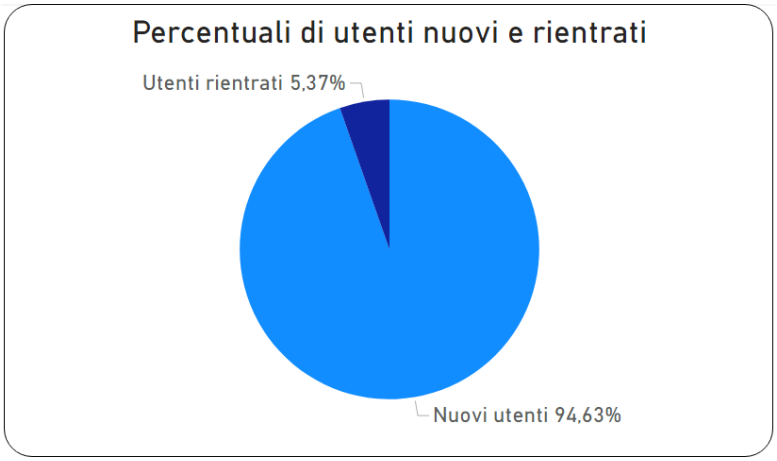


Figura 146.

Figura 147.

Dall'ulteriore seguente grafico, invece, si evince il numero di utenti per il medesimo tempo di stazionamento all'interno del sito. È stato determinato che i visitatori permangono principalmente per un arco di tempo che va da 0 a poco meno di 20 secondi, decrescendo di numero al trascorrere del tempo, come evidenziato dalla retta rappresentativa del trend.

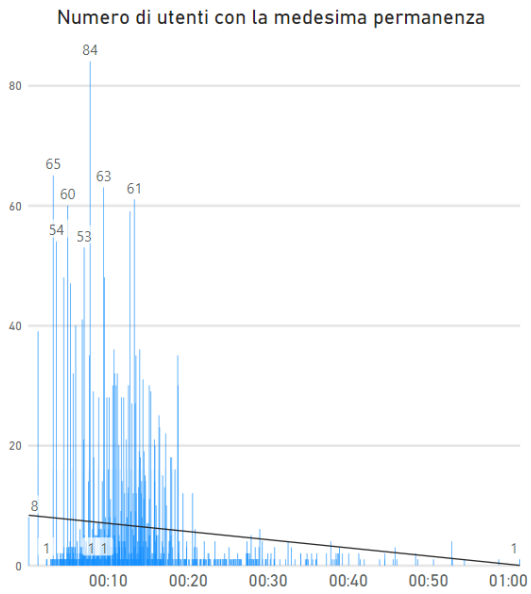


Figura 148.

Uno specifico report riporta il numero di utenti rientrati per il numero di accessi effettuati consentendo, pertanto, di constatare, in linea di massima, le volte che un utente deve rientrare per ritenersi soddisfatto, nel caso di potenziale interesse dei servizi

offerti dal sito. Dei 2532 utenti rientrati totali, poco meno della metà sono rientrati una seconda volta, più di 1/6 ben tre volte. Maggiore è il numero di ingressi effettuati e maggiore è presumibile l'interesse del cliente.

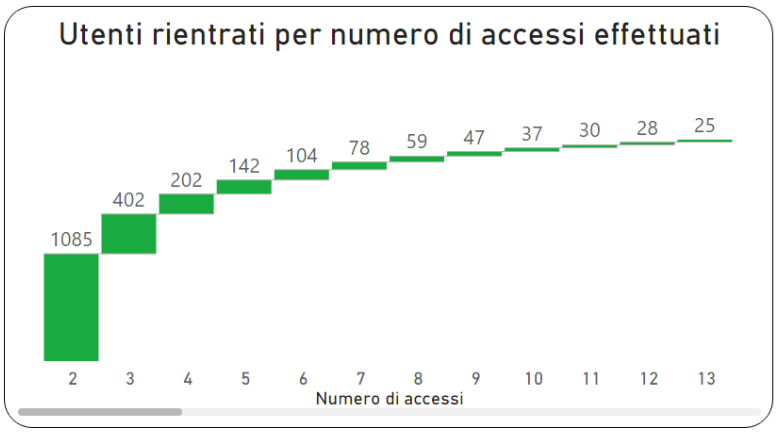


Figura 149.

Altra informazione desumibile dalla reportistica, ha riguardato il quantitativo di visitatori rientrati per giorni intercorsi dall'ultimo accesso. Dei 2532 rientrati, ben oltre la metà è rientrata il giorno stesso di avvenuto primo accesso, mentre solo 181 sono rientrati il giorno seguente: un numero esiguo rispetto al totale e irrilevante se confrontato ai circa 47000 utenti totali tra rientrati e non. Tali dati hanno confermato la necessità di improntare le direzioni successive verso un perfezionamento dell'appeal del sito, al fine di trattenere maggiormente gli utenti, sia per tempo di permanenza che per numero di accessi effettuati.

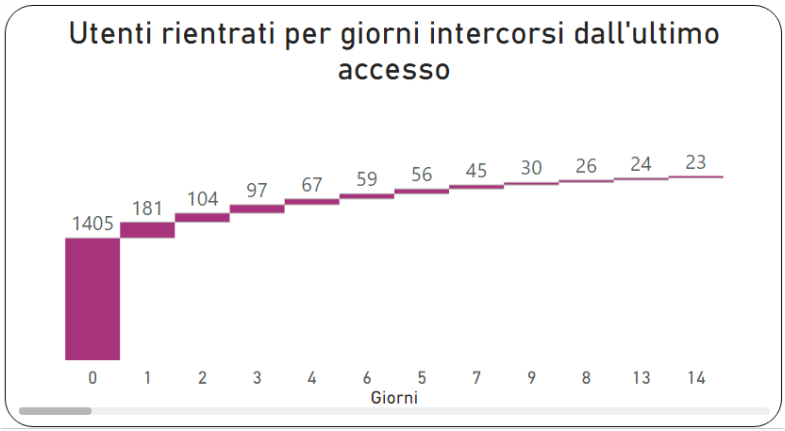


Figura 150.

3.8 Dashboard

Una volta realizzati i diversi report e formulato le differenti conclusioni e interpretazioni dei dati, si è passati alla fase di condivisione dell'elaborato. Ciò è stato possibile mediante l'ambiente dedito di Power BI Service che ha consentito il poter caricare rispettivamente il data set ed e le relative visuals. Da quest'ultime, si ha avuto modo di creare il dashboard, da porre in condivisione con l'eventuale resto del personale. È un servizio estremamente utile e comodo per poter meglio consentire l'allineamento e la comunicazione delle informazioni estrapolate. Si sono potuti creare più dashboard a seconda della tematica da voler nello specifico affrontare o dell'utente di destinazione ma, in questo caso, si è scelto di riportare per intero la reportistica su ogni aspetto affrontato.

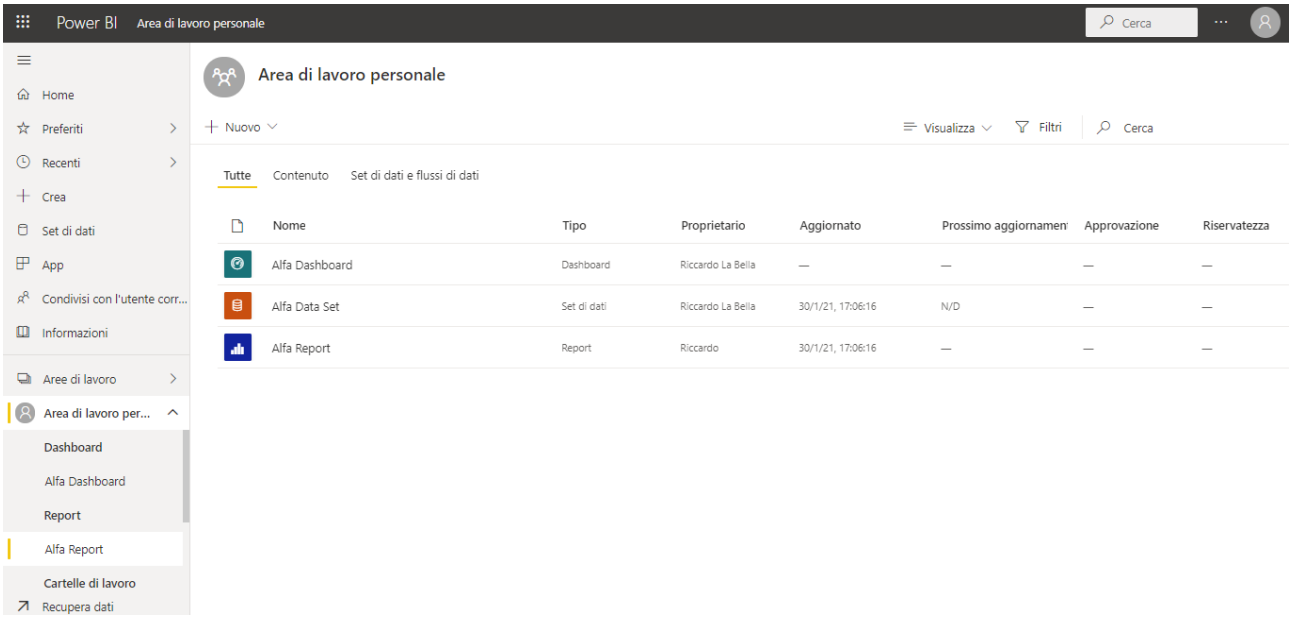


Figura 151.

CONCLUSIONI

Sulla scorta di quanto sopra espresso ed a conclusione dell'elaborato, si può affermare come la Business Intelligence, con i suoi processi tecnologici, che uniscono e combinano tra di loro analisi dei dati, strumenti, infrastrutture e best practice, sia in grado di orientare i decision makers nelle loro decisioni e strategie in modo puntuale ed efficace, ponendoli in grado di adottare le giuste decisioni a partire dai dati e dalle evidenze numeriche che essi esprimono; la Business Intelligence ha così assunto la funzione di strumento ormai indispensabile all'interno di una azienda, da implementare sempre più per trarre il massimo beneficio dai dati che raccoglie strada facendo. Solo la costante misurazione delle performance permette un'ottimizzazione dei processi, con conseguente risparmio di risorse e, solo un quadro chiaro e quanto mai più dettagliato dello stato vitale di un'azienda, e quindi di tutti i suoi processi interni, può consentire l'implementazioni di una strategia efficace, basata su informazioni approfondite, oggettive e rappresentative della realtà.

Nel prossimo futuro l'evoluzione della Business Intelligence sarà fortemente influenzato dal progresso tecnologico e dalla vieppiù disponibilità ed interconnessione dei dati reperibili al'interno, e non solo, della struttura aziendale. Se da un lato si potrà usufruire di più dati e tecniche più sofisticate per l'estrapolazione di dati da più e diverse sorgenti, dall'altro lato sarà possibile creare report informativi sempre più avanzati e particolareggiati, nel contempo più intuitivi e immediati grazie all'impiego di nuovi modelli e tecniche d'intelligenza artificiale.

Con riferimento al progetto previsto nel tirocinio, è stato possibile valutare, in una dinamica reale, le potenzialità e le opportunità che la BI offre in termini di produzione di report intelligibili, di facile analisi e interpretazione dei dati rielaborati, utili a individuare i punti di forza e di debolezza e di adottare, conseguentemente, gli opportuni interventi, tali da rendere maggiormente attrattivo il sito della società analizzato in maniera da indurre gli utenti ad un successivo rientro o, ancor meglio, assicurare sin dal primo

accesso del visitatore un efficace coinvolgimento ed un concreto interesse alle proposte di servizi ed alle sezioni visualizzabili.

Figure

Figura 1, *Logo Certimeter*, https://www.certimetergroup.com/

Figura 2, *Catena del valore*, Politecnico di Torino, Prof.ssa Alessandra Colombelli, Lucidi e materiale di Strategia e Organizzazione 2018/2019, https://didattica.polito.it/pls/portal30/sviluppo.pagina_studente_2016.main?t=2

Figura 3, *Piramide di Anthony*, Softeng, Politecnico di Torino, Prof. Fulvio Corno, Prof. Luca Ardito, Prof. Juan Pablo Saenz Moreno, Lucidi, materiale e dispense di Sistemi informativi 2018/2019, visitato il 22/11/20, https://elite.polito.it/teaching/current-courses/117-02cix?showall=1, Famiglie di SI, http://softeng.polito.it/courses/02CIXb/Dispense/cap-3-standalone.pdf

Figura 4*, Riepilogo*, Softeng, Politecnico di Torino, Prof. Fulvio Corno, Prof. Luca Ardito, Prof. Juan Pablo Saenz Moreno, Lucidi, materiale e dispense di Sistemi informativi 2018/2019, visitato il 22/11/20, https://elite.polito.it/teaching/current-courses/117-02cix?showall=1, Famiglie di SI, http://softeng.polito.it/courses/02CIXb/Dispense/cap-3-standalone.pdf

Figura 5, *Cubo*, Database anda Data Mining Group, Politecnico di Torino, Prof.ssa Laura Farinetti, Lucidi di Basi di Dati per Triennale di Ingegneria Gestionale, visitato il 4/12/20, https://dbdmg.polito.it/wordpress/teaching/basi-di-dati-gest/, Elena Baralis, Analisi dei dati per il supporto alle decisioni, https://dbdmg.polito.it/wordpress/wp-content/uploads/2019/12/SupportoDecisioni6x.pdf

Figura 6, *Dimensional Fact*, Database anda Data Mining Group, Politecnico di Torino, Prof.ssa Laura Farinetti, Lucidi di Basi di Dati per Triennale di Ingegneria Gestionale, visitato il 4/12/20, https://dbdmg.polito.it/wordpress/teaching/basi-di-dati-gest/, Elena Baralis, Analisi dei dati per il supporto alle decisioni, https://dbdmg.polito.it/wordpress/wp-content/uploads/2019/12/SupportoDecisioni6x.pdf

Figura 7, *Gerarchia*, Database anda Data Mining Group, Politecnico di Torino, Prof.ssa Laura Farinetti, Lucidi di Basi di Dati per Triennale di Ingegneria Gestionale, visitato il 4/12/20, https://dbdmg.polito.it/wordpress/teaching/basi-di-dati-gest/, Elena Baralis, Analisi dei dati per il supporto alle decisioni, https://dbdmg.polito.it/wordpress/wp-content/uploads/2019/12/SupportoDecisioni6x.pdf

Figura 8, *Diagramma ER*, Database anda Data Mining Group, Politecnico di Torino, Prof.ssa Laura Farinetti, Lucidi di Basi di Dati per Triennale di Ingegneria Gestionale, visitato il 4/12/20, https://dbdmg.polito.it/wordpress/teaching/basi-di-dati-gest/, Elena Baralis, Analisi dei dati per il supporto alle decisioni, https://dbdmg.polito.it/wordpress/wp-content/uploads/2019/12/SupportoDecisioni6x.pdf

Figura 9, *Gartner Magic Quadrant*, https://info.microsoft.com/ww-landing-2020-gartner-magic-quadrant-for-analytics-and-business-intelligence.html?LCID=EN-US

Figura 10, *Logo Power BI*, https://scholarbuys.com/power-bi-licensing-options/

Figura 11, *Struttura Power BI*, https://docs.microsoft.com/it-it/power-bi/fundamentals/power-bi-overview

Figura 12, *Funzionamento Power BI*, https://www.soluzioniedp.it/soluzioni/business-intelligence/microsoft-power-bi.html

Figura 13, Da Power BI Desktop.

Figura 14, Da Power BI Desktop.

Figura 15*, Loghi origine dati*, https://convverge.com/2018/01/31/power-bi-whats-big-deal/.

Figura 16 a 90*, Funzionalità*, da Power BI Desktop.

Figura 91 a 95, *Funzionalità*, Da Power BI Service.

Figura 96 a 99*, Data set*, Da Power BI Desktop.

Figura 100 a 102, Da Power BI Desktop.

Figura 103, *Gruppi*, Da Power BI Dektop.

Figura 104, *Delimitatore*, Da Power BI Desktop.

Figura 105, *Data Editor*, Da Power BI Desktop.

Figura 106, *Passaggi Applicati*, Da Power BI Desktop.

Figura 107*, Sostituisci i valori*, Da Power BI Desktop.

Figura 108, *Crea tabella*, Da Power BI Desktop.

Figura 109*, Data Editor*, Da Power BI Destop.

Figura 110, *Merge*, Da Power BI Desktop.

Figura 111, *Accoda*, Da Power BI Desktop.

Figura 112, *Merge,* Da Power BI Desktop.

Figura 113, *Data Editor*, Da Power BI Desktop.

Figura 114, *Aggiungi colonna condizionale*, Da Power BI Desktop.

Figura 115, *Data Editor*, Da Power BI Desktop.

Figura 116 a 120, *DAX*, Da Power BI Desktop.

Figura 121, *Data Model*, Da Power BI Desktop.

Figura 122, *Relazioni*, Da Power BI Desktop.

Figura 123, *Pagina Presentazione*, Da Power BI Desktop.

Figura 124, *Pagina Geografia*, Da Power BI Desktop.

Figura 125, *Pagina Visualizzazioni*, Da Power BI Desktop.

Figura 126, *Pagina Permanenza*, Da Power BI Desktop.

Figura 127, *Pagina Servizi*, Da Power BI Desktop.

Figura 128, *Pagina Utenti*, Da Power BI Desktop.

Figura 129*, Visualizzazioni per anno*, Da Power BI Desktop.

Figura 130, *Località di accesso*, Da Power BI Desktop.

Figura 131, *Percentuale traffico per pagina*, Da Power BI Desktop.

Figura 132, *Realtà italiana*, Da Power BI Desktop.

Figura 133, *Dati località,* Da Power BI Desktop.

Figura 134, *Regioni di provenienza degli accessi con interazioni*, Da Power BI Desktop.

Figura 135, *Dati pagine*, Da Power BI Desktop.

Figura 136, *Confronto dati visualizzazioni pagine 2019 e 2020*, Da Power BI Desktop.

Figura 137, *Confronto dati visualizzazioni per mese pagine 2019 e 2020*, Da Power BI Desktop.

Figura 138, *Pagina Visualizzazioni 2019*, Da Power BI Desktop.

Figura 139, *Pagina Visualizzazioni 2020*, Da Power BI Desktop.

Figura 140, *Confronto dati visualizzazioni per mese pagine 2018 e 2021*, Da Power BI Desktop.

Figura 141, *Dati visualizzazioni per mese pagine 2020*, Da Power BI Desktop.

Figura 142, *Confronto percentuali tra visualizzazioni e ore complessive per pagina da 2018 a 2021*, Da Power BI Desktop.

Figura 143, *Confronto ore e visualizzazioni per mese 2019 e 2020*, Da Power BI Desktop.

Figura 144, *Numero di visitatori nelle pagine dei servizi offerti*, Da Power BI Desktop.

Figura 145, *Distribuzione del tempo trascorso*, Da Power BI Desktop.

Figura 146, *Percentuale di utenti nuovi e rientrati*, Da Power BI Desktop.

Figura 147, *Utenti rientrati sul totale degli accessi*, Da Power BI Desktop.

Figura 148, *Numero di utenti con la medesima permanenza*, Da Power BI Desktop.

Figura 149, *Utenti rientrati per numero di accessi effettuati*, Da Power BI Desktop.

Figura 150, *Utenti rientrati per giorni intercorsi dall'ultimo accesso*, Da Power BI Desktop.

Figura 151, *Progetto,* Da Power BI Service.

Bibliografia

[6] Yansheng Chen, Zhijun Lin, *Business Intelligence Capabilities and Firm Performance: A Study in China,* International Journal of Information Management, Volume 57, Aprile 2021, 102232.

[9] João Lopes, Tiago Guimarães, Manuel Filipe Santos, *Adaptive Business Intelligence: A New Architectural Approach*, Procedia Computer Science, Volume 177, 2020, Pagine 540-545.

[10] Sandra Maria Correi, Loureiro, João Guerreiro, Iis Tussyadiah, *Artificial intelligence in business: State of the art and future research agenda*, Journal of Business Research, 20 Novembre 2020.

[11] Jaewoong Choi, Janghyeok Yoon, Jaemin Chung, Byoung-Youl Coh, Jae-MinLee*, Social media analytics and business intelligence research: A systematic review*, Information Processing & Management, Volume 57, Edizione 6, Novembre 2020, 102279.

[12] Vahid Khatibi, Abbas Keramati, Farid Shirazi, *Deployment of a business intelligence model to evaluate Iranian national higher education*, Social Sciences & Humanities Open, Volume 2, Edizione 1, 2020, 100056.

[13] Liang, Ting-Peng, Liu, Yu-His, *Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study,* Expert Systems with Applications, Volume 111, 30 Novembre 2018, Pagine 2-10.

[15] Harold Arturo Combita Niño, Johana Patricia Cómbita Niño, Roberto Morales Ortega, *Business intelligence governance framework in a university: Universidad de la costa case study*, International Journal of Information Management, Volume 50, Febbraio 2020, Pagine 405-412.

[24] Nuno Caseiro, Arnaldo Coelho, *The influence of Business Intelligence capacity, network learning and innovativeness on startups performance*, Journal of Innovation & Knowledge, Volume 4, Edizione 3, Luglio–Settembre 2019, Pagine 139-145.

[25] Valter Moreno, Flavia Cavazotte, Wagnerde Souza Carvalho, *Business intelligence and analytics as a driver of dynamic and operational capabilities in times of intense macroeconomic turbulence*, The Journal of High Technology Management Research, Volume 31, Edizione 2, Novembre 2020, 100389.

[26] Marilex Rea Llave, *Data lake in business intelligence: reporting from the trenches*, Procedia Computer Science, Volume 138, 2018, Pagine 516-524.

[27] Cong Cheng, Huihui Zhong, Liebing Cao, *Facilitating speed of internationalization: The roles of business intelligence and organizational agility*, Journal of Business Research, Volume 110, Marzo 2020, Pagine 95-103.

[28] Noor Ul Ain, Giovanni Vaia, William H.DeLone, Mehwish Waheed, *Two decades of research on business intelligence system adoption, utilization and success – A systematic literature review*, Decision Support Systems, Volume 125, Ottobre 2019, 113113.

[29] Katerina Božič, Vlado Dimovski, *Business intelligence and analytics use, innovation ambidexterity, and firm performance: A dynamic capabilities perspective*, The Journal of Strategic Information Systems, Volume 28, Edizione 4, Dicembre 2019, 101578.

[30] Pall Rikhardsson, Ogan Yigitbasioglu, *Business intelligence & analytics in management accounting research: Status and future focus, International Journal of Accounting Information Systems*, Volume 29, Giugno 2018, Pagine 37-58.

[31] Quoc Duy Vo, Jaya Thomas, Shinyoung Cho, Pradipta De, Bong Jun Choi, *Next Generation Business Intelligence and Analytics*, ICBIM '18: Proceedings of the 2nd International Conference on Business and Information Management, Settembre 2018, Pagine 163–168.

[34] Roger H. L. Chiang, Paulo Barcellos Góes, Edward A. Stohr, *Business Intelligence and Analytics Education, and Program Development: A Unique Opportunity for the Information Systems Discipline*, ACM Transactions on Management Information Systems, Ottobre 2012, Articolo No: 12.

[35] Craig Statchuk, *Business analytics trends and opportunities*, CASCON '11: Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research, Novembre 2011, Pagine 318–320.

[36] Bojan Cestnik, Stojan Košti*, Constructing semantic ontologies for business analytics*, CompSysTech '10: Proceedings of the 11th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing on International Conference on Computer Systems and Technologies, Giugno 2010, Pagine 436– 441.

[37] M. Golfarelli, Stefano Rizzi, *Data Warehouse: teoria e pratica della progettazione*, McGraw-Hill, seconda edizione, 2006.

[38] Qi Liu, Gengzhong Feng, Xi Zhao, Wenlong Wang, *Minimizing the data quality problem of information systems: A process-based method*, Decision Support Systems, Volume 137, Ottobre 2020, 113381.

[52] Thomas J. Dilling MD, *Artificial Intelligence Research: The Utility and Design of a Relational Database System*, Advances in Radiation Oncology, Volume 5, Edizione 6, Novembre-Dicembre 2020, Pagine 1280-1285.

[53] Benymol Jose, Sajimon Abraham, *Performance analysis of NoSQL and relational databases with MongoDB and MySQL*, Materials Today: Proceedings, Volume 24, Parte 3, 2020, Pages 2036-2043.

[54] Divya Mahajan, Cody Blakeney, Ziliang Zong, *Improving the energy efficiency of relational and NoSQL databases via query optimizations*, Sustainable Computing: Informatics and Systems, Volume 22, Giugno 2019, Pagine 120-133.

[55] Srdja Bjeladinovic, Zoran Marjanovic, Sladjan Babarogic, *A proposal of architecture for integration and uniform use of hybrid SQL/NoSQL database components*, Journal of Systems and Software, Volume 168, Ottobre 2020, 110633.

[70] Faizul Huq, Mahesh Kurpad, M.K.Raja, *The use of relational database management systems (DBMS) for information retrieval in a group technology (GT) environment*, Computers & Industrial Engineering, Volume 26, Edizione 2, Aprile 1994, Pagine 253-266.

[71] MingWang, H.Walker*, Creation of an intelligent process planning system within the relational DBMS software environment*, Computers in Industry, Volume 13, Edizione 3, Dicembre 1989, Pagine 215-228.

[72] C.J.Anumba, Data structures and DBMS for computer-aided design systems, Advances in Engineering Software, Volume 25, Edizione 2–3, Marzo–Aprile 1996, Pagine 123-129.

[73] Michael M.Gorman, *2 - DBMS APPLICATIONS AND COMPONENTS*, Database Management Systems, Understanding and Applying Database Technology, 1991, Pagine 31-51.

[74] T. William Olle, *Database management system (DBMS)*, Encyclopedia of Computer Science, Gennaio 2003, Pagine 517–520.

[75] G.Satyanarayana Reddy, Rallabandi Srinivasu, M.Poorna Chander, Srikanth Reddy Rikkula, *Data warehouse, data mining, olap and oltp technologies are essential elements to support decision-making process in industries*, (IJCSE) International Journal on Computer Science and Engineering, 2010, Volume 2, No. 09, Pagine 2865-2873.

[78] Hui Zhou, *Design of Student Information Management Database Application System for Office and Departmental Target Responsibility System*, Physics Procedia, Volume 25, 2012, Pagine 1660-1665.

[79] Gerald Post, Albert Kagan, *Database management systems: design considerations and attribute facilities*, Journal of Systems and Software, Volume 56, Volumr 2, 1 Marzo 2001, Pagine 183-193.

[80] Helland P. (2009) Database Management System. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi-org.ezproxy.biblio.polito.it/10.1007/978-0-387-39940-9_662.

[81] Judith Awiti, Alejandro A.Vaisman, Esteban Zimányi, *Design and implementation of ETL processes using BPMN and relational algebra*, Data & Knowledge Engineering, Volume 129, Settembre 2020, 101837.

[82] Cristiano Xavier, Fernando Moreira, *Agile ETL*, Procedia Technology, Volume 9, 2013, Pages 381-387.

[83] Panos Vassiliadis, Alkis Simitsis, Spiros Skiadopoulos, *Conceptual modeling for ETL processes*, Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP, Novembre 2002, Pagine 14–21.

[84] Tim A. Majchrzak, Tobias Jansen, Herbert Kuchen, *Efficiency evaluation of open source ETL tools*, Proceedings of the 2011 ACM Symposium on Applied Computing, Marzo 2011, Pagine 287–294.

[87] Robert R. Harmon, *Marketing Information Systems*, Encyclopedia of Information Systems, 2003, Pagine 137-151.

[88] Stephen R. Gardner, *Building the data warehouse*, Communications of the ACMSeptember, 1998.

[92] Zhengxin Chen, *Data Warehousing and Data Marts*, Encyclopedia of Information Systems, 2003, Pagine 521-533.

[93] Young-Koo Lee, Kyu-Young Whang, Yang-Sae Moon, Il-Yeol Song, *An aggregation algorithm using a multidimensional file in multidimensional OLAP*, Information Sciences, Volume 152, Giugno 2003, Pagine 121-138.

[94] David W.Cheung, Bo Zhou, Ben Kao, Hu Kan, Sau Dan Lee, *Towards the building of a dense-region-based OLAP system*, Data & Knowledge Engineering, Volume 36, Edizione 1, Gennaio 2001, Pagine 1-27.

[95] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining concepts and techniques*, Morgan Kaufmann Publishers is an imprint of Elsevier, Third Edition, 2012.

[96] Ana M Jimenez-Carvelo, Luis Cuadros-Rodríguez, *Data mining/machine learning methods in foodomics*, Current Opinion in Food Science, Volume 37, Febbraio 2021, Pagine 76-82.

[100] Matteo Golfarelli, Stefano Rizzi, Andrea Proli, *Designing what-if analysis: towards a methodology*, Proceedings of the 9th ACM international workshop on Data warehousing and OLAP, Novembre 2006, Pagine 51–58.

[101] Rasmussen, Nils H., *et al. Business Dashboards : A Visual Catalog for Design and Deployment*, John Wiley & Sons, Incorporated, 2009.

[102] Duarte de Carvalho, Ricardo Rocha, Vanessa Fernandes, Sofia Neves, *Business Intelligence: Future Perspectives (April, 2016)*, Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering, Giugno 2016.

[103] Darren Edge, Jonathan Larson, Christopher White, *Bringing AI to BI: Enabling Visual Analytics of Unstructured Data in a Modern Business Intelligence Platform*, Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, Aprile 2018 Paper No.: CS02, Pagine 1–9.

[108] Jim Lindell, MBA, CPA, CGMA, ANNUAL UPDATE FOR CONTROLLERS, Association of International Certified Professional Accountants & CIMA, 2020.

[114] Knight, Devin, *et al. Microsoft Power BI Quick Start Guide : Build Dashboards and Visualizations to Make Your Data Come to Life*, Packt Publishing, Limited, 2018.

[119] Senda Bouaziz, Ahlem Nablib, Faiez Gargouri, *Design a Data Warehouse Schema from Document-Oriented database*, Procedia Computer Science, Volume 159, 2019, Pagine 221-230.

Sitografia

[1] Università degli studi di Torino, *visitato il 11/01/21,* https://www.unito.it/search/google/certimeter#gsc.tab=0&gsc.q=certimeter&gsc.sort=

[2] Università degli studi di Torino, Dipartimento di informatica, Certimeter s.r.l. spin-off universitario, *visitato il 11/01/21,* https://www.unito.it/sites/default/files/scheda_certimeter.pdf

[3] Certimeter Group, Chi siamo e Servizi, *visitato il 11/01/21,* https://www.certimetergroup.com/

[4] Certimeter, Profilo, *visitato il 7/2/2021*, https://www.certimeter.it/Profilo.html

[5] Tableau, Business Intelligence, *visitato 17/11/20* il https://www.tableau.com/it-it/learn/articles/business-intelligence

[7] Extrasys, Che cos'è la Business Intelligence, *visitato il 17/11/20*, https://www.extrasys.it/it/redblog/che-cosa-%C3%A8-business-intelligence

[8] Wikipedia, Business Intelligence, *visitato il 17/11/20*, https://it.wikipedia.org/wiki/Business_intelligence

[14] Datamaze, Business Intelligence, *visitato il 18/11/20*, https://www.datamaze.it/risorse/tutto-sulla-business-intelligence

[16] Inside Agency, Storia Business Intelligence, *visitato il 26/11/20*, https://www.inside.agency/storia-business-intelligence/

[17] Slideshare, Storia Business Intelligence, *visitato il 26/11/20*, https://www.slideshare.net/lmauri/business-intelligence-v03

[18] Politecnico di Torino, Prof.ssa Alessandra Colombelli, Lucidi e materiale di Strategia e Organizzazione 2018/2019, https://didattica.polito.it/pls/portal30/sviluppo.pagina_studente_2016.main?t=2

[19] Wikipedia, Database Management System*, visitato il 17/11/20*, https://it.wikipedia.org/wiki/Database_management_system

[20] Digital4, Data-Driven, *visitato il 18/11/20,* https://www.digital4.biz/marketing/big-data-e-analytics/sei-regole-d-oro-per-un-data-driven-marketing-di-successo/

[21] Wikipedia, Data-Driven Marketing, *visitato il 18/11/20*, https://en.wikipedia.org/wiki/Data_driven_marketing

[22] Intensys, Business Intelligence e Business Analytics, *visitato il 18/11/20*, https://www.intesys.it/journal/digital-transformation/business-intelligence-e-analytics-la-creazione-di-valore-data-driven/

[23] Tableau, Business Intelligence e Business Analytics, *visitato il 22/11/20,* https://www.tableau.com/it-it/learn/articles/business-intelligence/bi-business-analytics

[32] Selecthub, Business Intelligence e Business Analytics, *visitato il 22/11/20*, https://www.selecthub.com/business-intelligence/business-intelligence-vs-business-analytics/

[33] E-service-online, Business Intelligence e Business Analytics, *visitato il 22/11/20*, https://www.e-service-online.com/business-analytics-ba-vs-business-intelligence-bi/

[39] Università di Bologna, Prof. Danilo Montesi, Lucidi sulla Business Intelligence, *visitato il 17/11/20*, http://www.cs.unibo.it/~montesi/CBD/07BI.pdf

[40] E-lite, Politecnico di Torino, Prof. Marco Torchiano, Prof. Fulvio Corno, Lucidi, materiale e dispense di Sistemi informativi 2018/2019*, visitato il 19/11/20*, https://elite.polito.it/teaching/current-courses/117-02cix?showall=1

[41] Capitolo 3, http://softeng.polito.it/courses/02CIXb/Dispense/cap-3-standalone.pdf

[42] Capitolo 10, https://elite.polito.it/files/courses/02CIX/libro/cap10-standalone.pdf

[43] Definizione sistema informativo, https://elite.polito.it/files/courses/02CIX/2018-2019/slide/01_IS_Definition.pdf

[44] Ingegneria dei requisiti, https://elite.polito.it/files/courses/02CIX/2018-2019/slide/04_Requirements.pdf

[45] Sistemi informativi manageriali, https://elite.polito.it/files/courses/02CIX/2018-2019/slide/08_KPI.pdf

[46] Softeng, Politecnico di Torino, Prof. Fulvio Corno, Prof. Luca Ardito, Prof. Juan Pablo Saenz Moreno, Lucidi, materiale e dispense di Sistemi informativi 2018/2019, *visitato il 22/11/20*,

http://softeng.polito.it/courses/02CIXa/#News

[47] Definizione di SI, http://softeng.polito.it/courses/02CIXa/podcast19/01-ISDefinition.pdf

[48] Famiglie di SI, http://softeng.polito.it/courses/02CIXb/Dispense/cap-3-standalone.pdf

[49] Database Master, Dati e Sistemi Informativi, *visitato il 22/11/20*, http://databasemaster.it/dati-e-software-in-azienda/

[50] Wikipedia, Sistema informativo aziendale, *visitato il 22/11/20*,

https://it.wikipedia.org/wiki/Sistema_informativo_aziendale

[51] Università di Bologna, Prof. Wilma Penzo, Lucidi di Sistemi informativi aziendali 2003/2004, http://www-db.deis.unibo.it/courses/SIL/LUCIDI/01_SIAziendali.pdf

[56] Oracle, Database, *visitato il 7/12/20*, https://www.oracle.com/it/database/what-is-database/#nosql

[57] Skuola.Net, Database distribuiti*, visitato il 7/12/20*, https://www.skuola.net/informatica/database.html

[58] Database anda Data Mining Group, Politecnico di Torino, Prof.ssa Laura Farinetti, Lucidi di Basi di Dati per Triennale di Ingegneria Gestionale, *visitato il 4/12/20*, https://dbdmg.polito.it/wordpress/teaching/basi-di-dati-gest/

[59] Elena Baralis, Introduzione alle basi di dati, 2007, https://dbdmg.polito.it/wordpress/wp-content/uploads/2011/10/1-IntroduzioneBD6x.pdf

[60] Database distribuiti e NoSQL, https://dbdmg.polito.it/wordpress/wp-content/uploads/2019/11/5.6-DatabaseDistribuiti-NoSQL6x.pdf

[61] Elena Baralis, Analisi dei dati per il supporto alle decisioni, https://dbdmg.polito.it/wordpress/wp-content/uploads/2019/12/SupportoDecisioni6x.pdf

[62] Database anda Data Mining Group, Politecnico di Torino, Prof.ssa Tania Cerquitelli, Lucidi di Business intelligence per Magistrale di Ingegneria Gestionale, *visitato il 4/12/20,* https://dbdmg.polito.it/wordpress/teaching/business-intelligence/

[63] Elena Baralis, Big data, https://dbdmg.polito.it/wordpress/wp-content/uploads/2018/10/1-DSIntro.pdf

[64] Elena Baralis, Progettazione di Datawarehouse, https://dbdmg.polito.it/wordpress/wp-content/uploads/2020/05/2-DWprog_x6pdf.pdf

[65] Elena Baralis, Analisi Olap, https://dbdmg.polito.it/wordpress/wp-content/uploads/2010/12/3-DWquery6.pdf

[66] Politecnico di Torino, Prof. Poncino Massimo, Lucidi di Basi di dati e sistemi informativi, 2004/2005, https://didattica.polito.it/pls/portal30/sviluppo.pagina_studente_2016.main?t=2

[67] Data Warehouse, https://file.didattica.polito.it/download/MATDID/11055606

[68] Introduzione ai database, https://file.didattica.polito.it/download/MATDID/01032309

[69] Sistemi informativi, https://file.didattica.polito.it/download/MATDID/21023647

[76] Università "G. D'annunzio" di Chieti-Pescara, Prof. Gianluca Amato, Lucidi di Economia informatica, https://www.sci.unich.it/~amato/teaching/old/datamining08/lucidi/03-olap.pdf

[77] HTML, Database e DBMS, *visitato il 4/12/20*, https://www.html.it/pag/31719/web-database-e-dbms/

[85] Ionos, Datawehouse*, visitato il 5/12/20*, https://www.ionos.it/digitalguide/online-marketing/analisi-web/dwh-che-cose-un-data-warehouse/

[86] Bismart, Datawarehouse, *visitato il 6/12/20*, https://blog.bismart.com/en/what-do-we-do-etl

[89] Università di Bologna, Prof. Stefano Rizzi, Lucidi Datawarehouse, http://www-db.disi.unibo.it/courses/SIG/DW1.pdf

[90] Fabiano Sileo, Data Warehouse, *visitato il 17/11/20,* https://www.fabianosileo.it/il-potere-dei-dati/cose-un-data-warehouse-e-perche-serve-alla-tua-azienda/

[91] Università della Calabria, Lucidi Datawarehouse, http://www.ecostat.unical.it/Didattica/Statistica/didattica/Metodi%20Statistici%20per%20le%20Aziende%201/10.Architetture%20di%20Data%20Warehouse.pdf

[97] Elena Baralis, Regole associative, http://dbdmg.polito.it/wordpress/wp-content/uploads/2010/12/6-DMasso6.pdf

[98] Elena Baralis, Classificazione, https://dbdmg.polito.it/wordpress/wp-content/uploads/2010/12/7-DMclass_x6.pdf

[99] Università di Bologna, Prof. Stefano Rizzi, What-if Analysis, http://www-db.deis.unibo.it/~srizzi/PDF/eds-WIA.pdf

[104] Zerounoweb, Business Intelligence e Analytics Platforms: ecco le migliori soluzioni secondo Gartner, *visitato il 25/1/21*, https://www.zerounoweb.it/analytics/business-intelligence-e-analytics-platforms-ecco-le-migliori-soluzioni-secondo-gartner/

[105] Gartner, *visitato il 25/1/21*, https://www.gartner.com/en

[106] Microsoft, Power BI, *visitato il 26/1/21*, https://docs.microsoft.com/it-it/learn/modules/introduction-power-bi/2-what-power-bi

[107] Microsoft, Power BI, visitato il 27/1/21, https://docs.microsoft.com/it-it/power-bi/fundamentals/power-bi-overview

[109] Wikipedia, Power BI, visitato il 28/1/21, https://it.wikipedia.org/wiki/Microsoft_Power_BI

[110] Microsoft, Power Pivot, *visitato il 17/11/20,* https://support.microsoft.com/it-it/office/power-pivot-potente-strumento-di-analisi-e-creazione-di-modelli-di-dati-in-excel-a9c2c6e2-cc49-4976-a7d7-40896795d045

[111] Eossolutions, Power Pivot, Power Query, Power BI, *visitato il 17/11/20*, https://www.eos-solutions.it/it/news-details/power-pivot-query-bi.html

[112] Excel Professionale, Power Pivot, Power Query, Power BI, *visitato il 17/11/20*, https://www.excelprofessionale.it/introduzione-a-power-pivot-e-power-bi/

[113] Facile BI, Power Pivot, Power Query, Power BI, *visitato il 17/11/20,* https://facilebi.it/power-bi/

[115] Netstrategy, Google Analytics, *visitato il 2/2/21*, https://www.netstrategy.it/web-marketing/google-analytics-cos%C3%A8-come-funziona-la-guida-completa

[116] Google, Google Analytics*, visitato il 2/2/21*, https://analytics.google.com/analytics/web/provision/?pli=1#/provision

[117] Google, Supporto Google Analytics, *visitato il 3/2/21*, https://support.google.com/analytics/answer/1257084#pageviews_vs_unique_views

[118] Google, Supporto Google Analytics, *visitato il 3/2/21, https://support.google.com/analytics/answer/1009409?hl=it*

[120] Wikipedia, ETL, *visitato il 17/11/20*, https://it.wikipedia.org/wiki/Extract,_transform,_load

[121] Wikipedia, Data Mining, *visitato il 17/11/20*, https://it.wikipedia.org/wiki/Data_mining

[122] Wikipedia, Olap*, visitato il 17/11/20*, https://it.wikipedia.org/wiki/OLAP

[123] Computer weekly, Olap e Data Mining, *visitato il 17/11/20,* https://www.computerweekly.com/news/2240024045/OLAP-and-data-mining-Whats-the-difference

[124] Classicfoxvalley, Olap e Data Mining, *visitato il 17/11/20*, https://it.classicfoxvalley.com/collate/difference-between-data-mining-and-olap/

[125] Database Master, Oltp e Olap, visitato il 22/11/20, http://databasemaster.it/dbms-oltp-dbms-olap/

[126] Ionos, Data Warehouse, *visitato il 17/11/20*, https://www.ionos.it/digitalguide/online-marketing/analisi-web/dwh-che-cose-un-data-warehouse/

[127] Inside Agency, Data Warehouse, *visitato il 17/11/20*, https://www.inside.agency/business-intelligence-data-warehousing-definizione-utilizzo/

[128] Amazon, Data Warehouse, *visitato il 17/11/20*, https://aws.amazon.com/it/data-warehouse/

[129] Microsoft, DAX*, visitato il 15/1/21*, https://docs.microsoft.com/it-it/dax/dax-overview