UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA
Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica e dell'Automazione



TESI DI LAUREA

Progettazione e realizzazione di una campagna di data science relativa all'impatto dell'Intelligenza Artificiale nella società

Design and implementation of a data science campaign exploring the impact of Artificial Intelligence on society

Relatore

Prof. Domenico Ursino

Candidato

Valentino Garbuglia

Correlatore

Dott. Michele Marchetti

ANNO ACCADEMICO 2025-2026

«Sapere è potere.»

Francesco Bacone

Sommario

La presente tesi affronta l'analisi dei dati relativi all'impatto dell'Intelligenza Artificiale nella società, con particolare riferimento al mercato del lavoro. L'obiettivo è fornire dapprima un punto di vista microscopico sulle carriere intraprendibili, le retribuzioni e i requisiti di assunzione, per poi passare a una prospettiva macroscopica analizzando lo sviluppo dell'AI in diversi Paesi. Il lavoro si focalizza sull'uso di Power BI per la pulizia e la visualizzazione di dataset pubblici ed eterogenei tramite lo sviluppo di dashboard orientate all'analisi descrittiva e diagnostica. È stata realizzata una fase completa di ETL (Extract, Transform, Load) per la trasformazione e l'integrazione di dati provenienti da fonti diverse. Fondamentale importanza è stata attribuita alla realizzazione di visualizzazioni interattive e alla scrittura di misure DAX per l'identificazione di relazioni e insight significativi. Il lavoro svolto dimostra come gli strumenti di Business Intelligence possano fornire un supporto concreto ai processi decisionali riguardanti le dinamiche e lo sviluppo del mercato del lavoro in ambito AI.

Parole chiave: Intelligenza Artificiale, Data Science, Data Analytics, Power BI, ETL, DAX, Data Visualization, Dashboard

Indice

Introduzione **1**

1 Introduzione alla Data Science **3**
1.1 Big Data . . . 3
1.2 Caratteristiche dei Big Data: le "5V" . . . 4
1.2.1 Volume . . . 4
1.2.2 Velocità . . . 5
1.2.3 Varietà . . . 5
1.2.4 Veracità . . . 7
1.2.5 Valore . . . 7
1.3 Data Science . . . 7
1.3.1 Ciclo di vita della Data Science: modello CRISP-DM . . . 8
1.4 Data Analytics . . . 10
1.4.1 Categorie di Data Analytics . . . 11
1.4.2 Differenze con la Data Analysis . . . 12
1.4.3 Ciclo di vita della Big Data Analytics . . . 12
1.5 Confronto tra Data Science e Data Analytics . . . 16

2 Introduzione a Power BI **17**
2.1 Power BI . . . 17
2.1.1 Vantaggi e limiti . . . 17
2.1.2 Modalità di utilizzo . . . 19
2.2 Componenti . . . 20
2.2.1 Architettura . . . 20
2.2.2 Operatività . . . 22
2.3 Data Visualization . . . 25
2.3.1 Vista Report . . . 26
2.3.2 Gestione dei Filtri . . . 28
2.3.3 Vista Dati . . . 28
2.3.4 Vista Relazioni . . . 29
2.4 Linguaggi di programmazione . . . 29

3 Descrizione dei dataset di partenza **31**
3.1 Dataset utilizzati . . . 31
3.1.1 Dataset 1: Global AI Job Market and Salary Trends 2025 . . . 32

3.1.2 Dataset 2: AI Global Index . . . 33
3.2 Processo di ETL . . . 33
3.2.1 Introduzione . . . 34
3.2.2 Attività di ETL effettuate sul primo dataset . . . 35
3.2.3 Attività di ETL effettuate sul secondo dataset . . . 42

4 Progettazione e realizzazione delle dashboard di base **46**
4.1 Panoramica dei visual e grafici disponibili in Power BI . . . 46
4.2 Misure DAX implementate . . . 48
4.3 Descrizione e analisi delle dashboard realizzate . . . 55
4.3.1 Dashboard 1: Global Market and Remote Landscape . . . 55
4.3.2 Dashboard 2: Skills and Compensation Analysis . . . 57
4.3.3 Dashboard 3: Skills and Educational Mapping . . . 58
4.3.4 Dashboard 4: Market Pulse and Hiring Dynamic . . . 59

5 Progettazione e realizzazione delle dashboard complesse **62**
5.1 Visual e grafici avanzati . . . 62
5.2 Descrizione e analisi delle dashboard realizzate . . . 63
5.2.1 Dashboard 5: Strategic Pillars and Global Positioning . . . 64
5.2.2 Dashboard 6: Global Salaries and Cluster Performance . . . 65
5.2.3 Dashboard 7: Predictive Career and Market Growth . . . 69
5.2.4 Dashboard 8: Market Dynamics and AI Readiness Landscape . . . 71

6 Discussione **73**
6.1 Punti di forza dell'analisi condotta . . . 73
6.1.1 Integrazione dei dati . . . 73
6.1.2 Processo di ETL . . . 73
6.1.3 Uso di Power BI . . . 73
6.1.4 Analisi di dashboard descrittive e diagnostiche . . . 74
6.2 Limiti e vulnerabilità dell'analisi condotta . . . 74
6.2.1 Qualità dei dati . . . 74
6.2.2 Difficoltà di allineamento dei dataset . . . 74
6.2.3 Limiti temporali e copertura geografica dei dataset . . . 74

Conclusioni e sviluppi futuri **75**

Bibliografia **76**

Sitografia **77**

Ringraziamenti **79**

Elenco delle figure

1.1 Caratteristiche dei Big Data 4
1.2 Differenze tra dati strutturati e non strutturati 6
1.3 Il diagramma di Venn della Data Science ideato da Drew Conway 8
1.4 Modello CRISP-DM 10
1.5 Categorie di Data Analytics 12
1.6 Ciclo di vita Big Data Analytics 15

2.1 Architettura di Power BI 20
2.2 Connessione ai dati 23
2.3 Power Query Editor 24
2.4 Pannello di visualizzazioni 24
2.5 Relazione tra dashboard e report 25
2.6 Comando di condivisione dei report 25
2.7 Tre tipi di visualizzazione in Power BI Desktop: Vista Report, Vista Dati, Vista Relazioni 26
2.8 Vista Report: pannello Filtri, Visualizzazioni e Campi 27
2.9 Esempio di Vista Dati 28
2.10 Esempio di Vista Relazioni 29

3.1 Fasi del processo di ETL 35
3.2 Importazione dei dati in Power BI 36
3.3 Anteprima del dataset da importare e le opzioni di caricamento 36
3.4 Vista delle relazioni del modello analizzato 45

4.1 Global Market and Remote Landscape 56
4.2 Skills and Compensation Analysis 57
4.3 Skills and Educational Mapping 59
4.4 Market Pulse and Hiring Dynamic 60

5.1 Strategic Pillars and Global Positioning 64
5.2 Global Salaries and Cluster Performance 66
5.3 Predictive Career and Market Growth 69
5.4 Market Dynamics and AI Readiness Landscape 71

Elenco delle tabelle

1.1 Confronto tra Data Science e Data Analytics 16

3.1 Mappatura dei codici della colonna *experience_level* 38
3.2 Mappatura dei codici della colonna *employment_type* 38

Listings

3.1 Modifica del tipo di dato nella colonna *benefits_score* . . . 37
3.2 Sostituzione di dati numerici mancanti nel primo dataset . . . 37
3.3 Sostituzione di dati testuali mancanti nel primo dataset . . . 37
3.4 Sostituzione dei codici della colonna *experience_level* . . . 38
3.5 Sostituzione dei codici della colonna *employment_type* . . . 39
3.6 Normalizzazione dei testi nelle colonne *job_title* ed *education_required* . . . 39
3.7 Rimozione della colonna *required_skills* . . . 39
3.8 Aggiunta della colonna *days_to_apply* . . . 40
3.9 Ridenominazione delle intestazioni delle colonne . . . 40
3.10 Aggiunta della colonna *Order of Experience Level* . . . 41
3.11 Pulizia della colonna *Company Location* . . . 41
3.12 Aggiunta della colonna *Work Mode* . . . 41
3.13 Operazioni sulla tabella *Skill_Analysis* . . . 42
3.14 Modifica dei tipi di dati nel secondo dataset . . . 43
3.15 Sostituzione di dati numerici mancanti nel secondo dataset . . . 43
3.16 Sostituzione di dati testuali mancanti nel secondo dataset . . . 43
3.17 Modifica dei nomi dei Paesi nel secondo dataset . . . 44
3.18 Operazione di Unpivot sulle colonne selezionate . . . 44
4.1 Misura DAX di *% Market Penetration* . . . 48
4.2 Misura DAX di *Average Salary* . . . 49
4.3 Misura DAX di *Average Years Experience* . . . 49
4.4 Misura DAX di *Counting of Skills* . . . 49
4.5 Misura DAX di *Full-Remote Percentage* . . . 49
4.6 Misura DAX di *Global Average Total Score* . . . 49
4.7 Misura DAX di *Global Benchmark* . . . 50
4.8 Misura DAX di *Hiring Velocity* . . . 50
4.9 Misura DAX di *Key Strength Display* . . . 50
4.10 Misura DAX di *New Job Openings (30d)* . . . 50
4.11 Misura DAX di *Normalized Salary* . . . 51
4.12 Misura DAX di *Overall Score Display* . . . 51
4.13 Misura DAX di *Peak Salary* . . . 51
4.14 Misura DAX di *Remote Adoption Rate* . . . 51
4.15 Misura DAX di *Salary Gap* . . . 52
4.16 Misura DAX di *Strategic Gap* . . . 52
4.17 Misura DAX di *Target Benchmark* . . . 52

4.18 Misura DAX di *Top Hiring Country* 52
4.19 Misura DAX di *Total Job Postings* 53
4.20 Misura DAX di *Weakest Pillar (Dormant)* 53
4.21 Misura DAX di *World Rank Display* 54
4.22 Colonna calcolata *Market Group* 54
4.23 Colonna calcolata *Month_EN* 54
4.24 Colonna calcolata *Sort Month Date* 54
4.25 Colonna calcolata *Year_Filter* 54
5.1 Script Python per la realizzazione della Bubble Matrix 66
5.2 Script Python per la realizzazione dell'Hexbin Plot 67
5.3 Script Python per la realizzazione della Correlation Matrix 69

Introduzione

Negli ultimi anni l'Intelligenza Artificiale si sta diffondendo all'interno della società e nel contesto economico globale, trasformando in modo radicale il modo di lavorare in quasi tutti i settori. L'evoluzione rapida dei modelli di machine learning e delle tecnologie generative ha aperto un dibattito molto acceso sulle conseguenze reali che questa innovazione avrà sull'occupazione.

Più che a una semplice eliminazione dei ruoli tradizionali, si assiste a una profonda ridefinizione dei contesti lavorativi, in cui l'automazione delle attività operative sposta il focus verso nuove competenze tecniche, influenzando sia le dinamiche salariali sia i requisiti richiesti dalle aziende.

In questo scenario, la capacità di raccogliere, elaborare e interpretare grandi volumi di dati rappresenta uno degli strumenti più efficaci per comprendere l'evoluzione del mercato del lavoro e guidare le strategie occupazionali. Di conseguenza, è emersa in modo evidente la necessità di adottare approcci analitici in grado di fornire una visione approfondita dei fenomeni legati al mondo del lavoro in ambito AI.

Il lavoro svolto in questa tesi nasce proprio con l'obiettivo di analizzare l'impatto dell'Intelligenza Artificiale sul mondo del lavoro seguendo due direttive, ovvero una prospettiva di dettaglio (microscopica), incentrata sulle singole carriere, sui livelli salariali e sui requisiti di assunzione, e una visione d'insieme (macroscopica), orientata a confrontare la diffusione e lo sviluppo dell'AI tra diversi Paesi.

La scelta di utilizzare Power BI, una piattaforma largamente diffusa e accessibile, si basa sulla necessità di dimostrare come strumenti di Business Intelligence permettano di trasformare dati complessi in dashboard interattive e di facile lettura, utili anche per chi deve prendere decisioni strategiche senza essere un esperto nel settore.

La piattaforma Kaggle rappresenta un punto di riferimento fondamentale in questo ambito, mettendo a disposizione un'ampia raccolta di dataset pubblici reali, eterogenei e complessi utili a esplorare le dinamiche del settore.

Nel corso della tesi adottiamo un approccio diviso in più fasi. Inizialmente scegliamo due dataset contenenti dati sulle offerte di lavoro e le relative informazioni pubblicate da diversi Paesi in un determinato periodo di tempo, oltre a indici che rappresentano lo sviluppo dell'AI in varie nazioni. Successivamente sottoponiamo questi dati a un processo di ETL (Extract, Transform, Load) finalizzato alla pulizia, trasformazione e integrazione dei dati provenienti da fonti eterogenee, prestando particolare attenzione alla qualità e alla coerenza delle informazioni.

In seguito, ci concentriamo sulla fase analitica, eseguendo un'analisi di tipo descrittivo per comprendere la distribuzione e le caratteristiche dei fenomeni osservati, e un'analisi

diagnostica volta a esplorare relazioni e correlazioni tra le variabili in esame. Per eseguire questa fase di analisi realizziamo delle dashboard interattive in Power BI sfruttando le colonne presenti nei dataset e le misure DAX implementate, progettate per favorire una visualizzazione intuitiva dei dati e una corretta interpretazione dei risultati. Le dashboard sviluppate sono suddivise in due gruppi: un primo gruppo relativo alle dashboard di "base", in cui sono presenti solo grafici standard disponibili in Power BI, e un secondo gruppo, comprendente le dashboard "complesse", in cui impieghiamo visual e grafici avanzati, sia importati dallo store di Power BI sia realizzati con script Python. La principale differenza tra questi due gruppi risiede nella loro capacità espressiva: le dashboard di base risultano più adatte a un'analisi descrittiva, mentre quelle avanzate consentono di individuare relazioni non immediatamente visibili tra i dati, rivelandosi particolarmente efficaci per l'analisi diagnostica.

Il contributo della tesi consiste nel dimostrare come un approccio strutturato di Data Science, focalizzato sulla fase analitica e supportato da strumenti di Business Intelligence, consenta di trasformare dati eterogenei in informazioni utili per analizzare nel dettaglio l'impatto dell'Intelligenza Artificiale, individuando correlazioni significative e supportando le strategie decisionali.

Questa tesi è strutturata come di seguito specificato:

- Il Capitolo 1 introduce il concetto di Data Science a partire dai Big Data, illustrandone le caratteristiche principali (5V), per poi introdurre la Data Analytics e proporre un confronto tra i due ambiti.

- Il Capitolo 2 presenta Power BI, la piattaforma di Business Intelligence utilizzata per lo sviluppo del presente lavoro, descrivendone le funzionalità, le componenti, i vantaggi, i limiti applicativi legati all'analisi dei dati e il suo utilizzo per la Data Visualization.

- Il Capitolo 3 descrive i dataset utilizzati per l'analisi, illustrando le operazioni di selezione, pulizia e integrazione (Extract, Transform, Load - ETL), necessarie alla preparazione dei dati per l'analisi.

- Il Capitolo 4 illustra i grafici standard di Power BI e le misure DAX implementate, per poi analizzare nel dettaglio le prime quattro dashboard di base.

- Il Capitolo 5 presenta i visual avanzati impiegati e le restanti quattro dashboard complesse.

- Il Capitolo 6 propone una riflessione critica sul lavoro svolto, valutando i diversi punti di forza, le vulnerabilità e i limiti.

CAPITOLO 1

Introduzione alla Data Science

In questo capitolo viene introdotto il concetto di Data Science a partire dai Big Data, di cui si illustrano le principali caratteristiche, il loro valore e i benefici che hanno portato alle diverse organizzazioni. In particolare, vengono descritte le "5V" dei Big Data (volume, velocità, varietà, veracità e valore), che ne rappresentano le caratteristiche fondamentali. Successivamente, vengono approfonditi nel dettaglio la Data Science e il suo ciclo di vita secondo il modello CRISP-DM. Viene, poi, presentata la Data Analytics con le sue diverse tipologie di analisi, tracciando un confronto con la Data Analysis e descrivendo il ciclo di vita della Big Data Analytics. Il capitolo si conclude con la Data Science e la Data Analytics che vengono messe a confronto per evidenziarne le differenze, ma anche per comprendere come i loro ruoli nei processi di elaborazione dei dati siano complementari.

1.1 Big Data

I Big Data si riferiscono a set di dati estremamente grandi e complessi che non possono essere facilmente gestiti o analizzati con gli strumenti di elaborazione tradizionali. Per contenere i Big Data si utilizzano sistemi con elevata scalabilità e soluzioni basate su NoSQL (ossia sistemi software dove la persistenza dei dati è caratterizzata dal fatto di non utilizzare il modello relazionale normalmente adottato dagli RDBMS). Questi archivi di dati, di norma, non richiedono uno schema fisso (schemaless), possono evitare le operazioni di unione e puntano a scalare in modo orizzontale.

La quantità e la disponibilità di dati stanno crescendo rapidamente, grazie ai progressi della tecnologia digitale, come la connettività, la mobilità, l'Internet of Things (IoT) e l'Intelligenza Artificiale (AI). Man mano che i dati continuano a espandersi e a proliferare, emergono nuovi strumenti per big data per aiutare le aziende a raccoglierli, elaborarli e analizzarli alla velocità necessaria.

Il valore dei Big Data risiede nella capacità di essere trasformati in informazioni utili, tramite analisi avanzate, per migliorare i processi decisionali aziendali. Tuttavia, nella loro forma originaria, essi si presentano allo stato grezzo: spesso sono ridondanti, poco utili e con elevata presenza di "rumore". Senza un approccio intelligente, questa mole informativa rischia di rimanere inutilizzata. Per questo motivo, l'applicazione di tecniche avanzate di analytics e di operazioni di data cleaning permette di arrivare al concetto di *Smart Data*. Gli Smart Data rappresentano la componente dei Big Data a cui viene attribuita una precisa semantica, estraendo, così, il reale patrimonio informativo utile all'azienda o al processo analizzato.

L'adozione di un approccio orientato agli Smart Data permette alle organizzazioni diversi benefici, tra cui:

- il miglioramento del processo decisionale;
- una maggiore agilità e innovazione;
- l'ottimizzazione delle esperienze per i clienti;
- la possibilità di effettuare analisi e monitoraggi in tempo reale;
- l'incremento dell'efficienza delle operazioni interne;
- una gestione dei rischi migliorata.

1.2 Caratteristiche dei Big Data: le "5V"

Le principali caratteristiche dei Big Data si possono riassumere nel modello delle "5V", che definiscono la gestione e l'analisi di enormi quantità di informazioni. Nel 2001, Doug Laney, allora vicepresidente e Service Director dell'azienda Meta Group, descrisse in un report il Modello delle "3V" relativo alle tre dimensioni fondamentali dei Big Data: volume, velocità e varietà. Un modello semplice e sintetico per definire dei nuovi dati, generati dall'aumento delle fonti informative e, più in generale, dall'evoluzione delle tecnologie. Il paradigma di Laney è stato poi arricchito dalle variabili di veracità e valore; per questo si parla di "5V" dei Big Data (Figura 1.1).

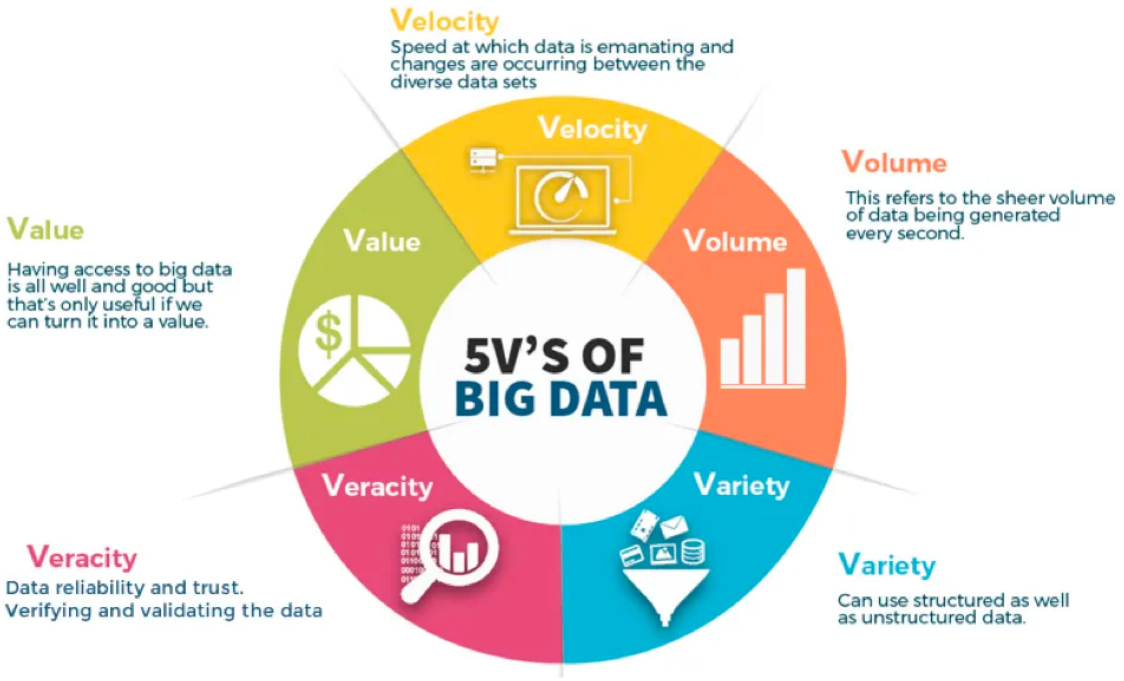


Figura 1.1: Caratteristiche dei Big Data

1.2.1 Volume

Il volume rappresenta la mole di dati generati e raccolti. La quantità di dati è esplosa negli ultimi anni, grazie alla digitalizzazione crescente e all'espansione di Internet. Le principali fonti di questi dati sono:

- i dispositivi IoT e la sensoristica, ovvero miliardi di endpoint che generano flussi continui di dati in tempo reale;
- i log di sistema, ovvero file di registro generati da infrastrutture IT e macchine, essenziali per attività di Log Analytics;
- le applicazioni enterprise, come software transazionali aziendali, sistemi ERP e piattaforme di CRM;
- le attività umane, tra cui interazioni sui social media, transazioni finanziarie, dati di clickstream e cartelle cliniche digitali;
- gli Open Data e la ricerca, ovvero dataset condivisi dalla comunità scientifica e da enti istituzionali.

I cosiddetti "grandi volumi di dati" si riferiscono, quindi, non solo a terabyte o petabyte, ma anche a enormi dataset in costante espansione, prodotti senza interruzione e difficili da contenere nei sistemi tradizionali. Le aziende che operano in settori come manifatturiero, retail, energy o finanza raccolgono ogni giorno dati che richiedono strumenti di archiviazione ed elaborazione scalabili, progettati per gestire dati eterogenei e frequenti aggiornamenti.

Per trattare con efficacia tali quantità, servono architetture moderne, come Data Lake e Data Lakehouse, basate su tecnologie cloud-native (Databricks, Microsoft Fabric). Queste piattaforme consentono di centralizzare dati massivi, mantenere performance elevate e creare la single source of truth che alimenta gli use case avanzati di Business Intelligence e AI predittiva.

1.2.2 Velocità

La velocità riflette la rapidità con cui i dati vengono prodotti, trasmessi e resi disponibili per l'analisi. In molti contesti aziendali, specialmente legati al mondo industriale, i flussi informativi devono essere elaborati in tempo reale o near-real-time per generare valore.

Per gestire questa dinamicità servono tecnologie progettate per lo streaming e l'elaborazione incrementale: event processing, pipeline in streaming, sistemi di Complex Event Processing (CEP), database capaci di memorizzare serie temporali e architetture orientate ai microservizi. Tali strumenti consentono di individuare anomalie operative, reagire a eventi critici, aggiornare previsioni e modelli predittivi con frequenza molto elevata.

Dal punto di vista architetturale la gestione della velocità impone il superamento dei tradizionali modelli di elaborazione di tipo batch. Con questo modello, i sistemi eseguono una serie di programmi che accettano un insieme di file di dati come input, elaborano i dati e producono un insieme di file di dati come output. Nel modello batch vengono elaborati blocchi di dati che sono già stati memorizzati in un determinato periodo di tempo. È così chiamato perché i dati vengono raccolti in batch come set di record ed elaborati come un'unità. Al contrario, l'elaborazione stream è un metodo utilizzato per eseguire query su flussi continui di dati e rilevare rapidamente le condizioni entro un periodo di tempo limitato. In altre parole, l'elaborazione stream è l'elaborazione dei dati direttamente mentre questi vengono prodotti o ricevuti. Per ottenere valore dai big data, i dati devono essere elaborati non appena arrivano e tale elaborazione deve essere di qualità.

1.2.3 Varietà

La varietà dei dati si riferisce a formati multipli e tipi di dati eterogenei, che devono essere supportati da soluzioni di Big Data. Questa varietà comporta sfide per le imprese in termini

di integrazione, trasformazione, elaborazione e memorizzazione dei dati. I tre principali tipi di dati sono:

- *Dati strutturati:* sono organizzati secondo uno schema rigido e un modello di gestione relazionale, dunque ben definito come un database composto da tabelle, fogli di calcolo o rapporti statistici. Tale modello è semplice da interrogare. Questi dati vengono estratti, interpretati e memorizzati in un archivio formattato, come un database SQL. In questo tipo di struttura vi sono svariati elementi, come tabelle, righe e colonne e chiavi relazionali, che permettono di mappare i dati in campi predefiniti.
- *Dati non strutturati:* sono presenti in diversi formati e in una forma grezza: non hanno uno schema predefinito né un formato standard, e possono essere file di testo, file multimediali, come immagini, video, audio, documenti in PDF. A differenza dei dati strutturati, i dati non strutturati non sono adatti a un database relazionale e sono più complessi da analizzare. Infatti richiedono piattaforme alternative e tecniche di elaborazione più avanzate che riescano a individuare e interpretare il contenuto dei dati senza uno schema predefinito. Gli strumenti di elaborazione del linguaggio naturale (NLP) aiutano a comprendere i dati non strutturati che esistono in formato scritto.
- *Dati semi-strutturati:* come, per esempio, i file JSON o XML; essi hanno alcune regole di formattazione ma una struttura più flessibile e stanno a metà tra dati strutturati e non strutturati. Questo genere di dati contiene informazioni con caratteristiche ibride, con una struttura flessibile, ma non "costretti" in un database relazionale. Ciò nonostante, i dati semi-strutturati sono caratterizzati da alcune proprietà organizzative che ne facilitano l'analisi, ossia contengono anche informazioni aggiuntive, come metadati o tag, che li rendono più organizzati rispetto ai dati non strutturati.

Nella Figura 1.2 vengono specificate le principali differenze tra dati strutturati e non strutturati.

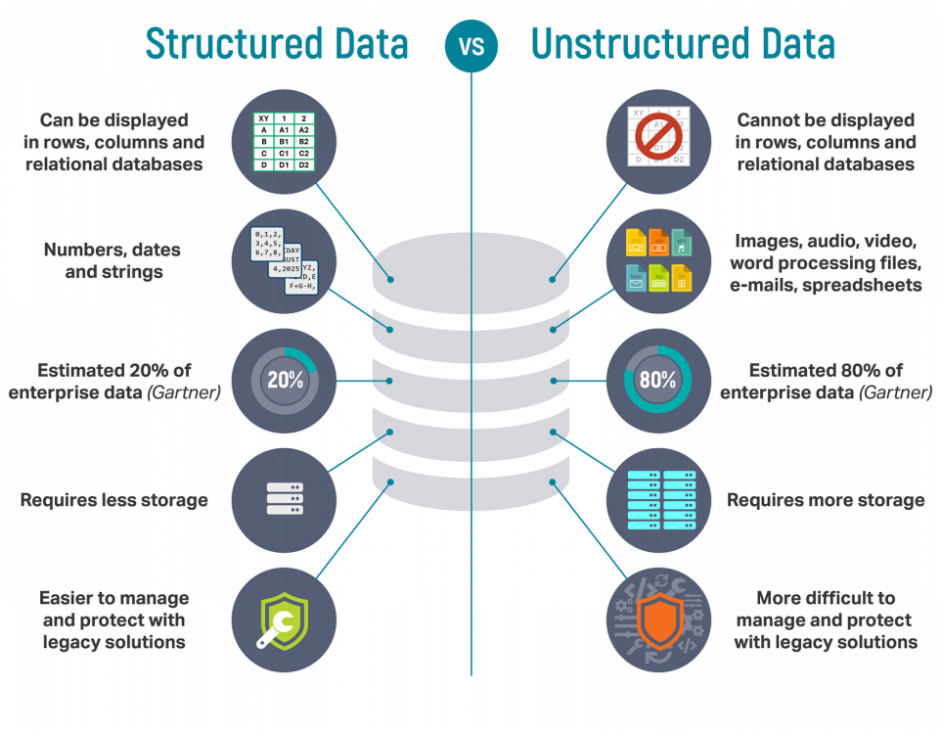


Figura 1.2: Differenze tra dati strutturati e non strutturati

1.2.4 Veracità

La veracità nei Big Data si riferisce alla qualità, alla coerenza e all'affidabilità dei dati. A differenza dei sistemi informativi tradizionali, in cui la qualità del dato è generalmente garantita dalla presenza di rigidi vincoli di integrità e processi strutturati di validazione, i Big Data sono caratterizzati da una grande eterogeneità. La diversa natura delle fonti introduce inevitabilmente elementi di incertezza, rumore e bias informativo. La conseguenza di ciò è che la qualità dei dati può variare, rendendo i processi decisionali vulnerabili al rischio di conclusioni distorte o errate. Diventa, quindi, cruciale utilizzare tecniche di Data Cleaning per mitigare le anomalie dei dati, preservandone l'accuratezza analitica e garantendo che gli insight estratti poggino su una base informativa solida e affidabile. Disporre di una mole immensa di dati (volume) ad alta velocità (velocità) diventa controproducente, se il livello di fiducia riposto in essi è basso; questo fenomeno viene definito *paradosso della quantità*. La veracità diventa, quindi, fondamentale per trasformare il dato grezzo in Smart Data, ovvero in uno strumento strategico capace di generare reale valore decisionale per le organizzazioni.

1.2.5 Valore

Il valore è l'obiettivo finale del processo di analisi dei Big Data, poiché risiede nella capacità di estrarre preziose informazioni per orientare il processo decisionale, ottimizzare le operazioni e creare opportunità. L'implementazione di queste tecnologie consente alle organizzazioni di convertire dati grezzi e non elaborati in risultati fruibili, capaci di migliorare le customer experience e l'efficienza operativa, oppure di promuovere innovazioni in settori come l'assistenza sanitaria e il retail.

Il valore è strettamente correlato alla veracità: se l'analisi si basa su informazioni affette da rumore o bias, il valore estratto risulta nullo o fuorviante; il valore emerge, quindi, solo quando il volume e la velocità dei dati vengono filtrati da rigidi processi di pulizia e validazione. Un ulteriore fattore determinante che può influenzare questa dimensione è il tempo. Il potenziale strategico dei dati non è statico, ma tende a deperire con il trascorrere del tempo; di conseguenza la tempestività dell'analisi permette di trasformare un'informazione grezza in un reale vantaggio competitivo.

1.3 Data Science

La Data Science è lo studio scientifico dei dati finalizzato ad ottenere conoscenza. Questo campo combina più discipline per estrarre informazioni da set di dati di grandi dimensioni con l'obbiettivo di supportare processi decisionali e di formulare modelli predittivi. Scienziati dei dati, analisti dei dati, data architect, ingegneri dei dati, statistici, amministratori di database e analisti aziendali lavorano tutti nel campo della Data Science. L'importanza della Data Science è in costante crescita, guidata dall'espansione esponenziale dei dati disponibili e dalla necessità, da parte delle organizzazioni, di adottare un approccio data-driven per stimolare l'innovazione e ottimizzare le performance competitive. La conoscenza in questo ambito scaturisce dall'intersezione di tre competenze fondamentali:

- *conoscenza matematica e statistica*: l'uso di modelli matematici, formule ed equazioni per l'analisi quantitativa e la validazione dei dati;
- *capacità informatiche*: l'utilizzo di linguaggi di programmazione e architetture software per l'elaborazione efficiente dei dataset;
- *esperienza o conoscenza del dominio*: la conoscenza del contesto applicativo del problema (ad esempio medicina, finanza, scienze sociali, ecc.).

Come viene evidenziato dallo schema di Conway (Figura 1.3), la Data Science non si limita alla semplice somma di queste abilità, ma ad una loro vera e propria sinergia. Dall'unione delle capacità informatiche e matematiche si dà vita al *machine learning*. Il grafico però mette in guardia da una *zona pericolosa*, in cui l'applicazione di tecniche di programmazione a un problema specifico, se priva del necessario rigore statistico, rischia di produrre conclusioni fuorvianti. La corretta combinazione di questi elementi permette di governare la complessità dei dati, un aspetto che richiede un approccio metodologico rigoroso.

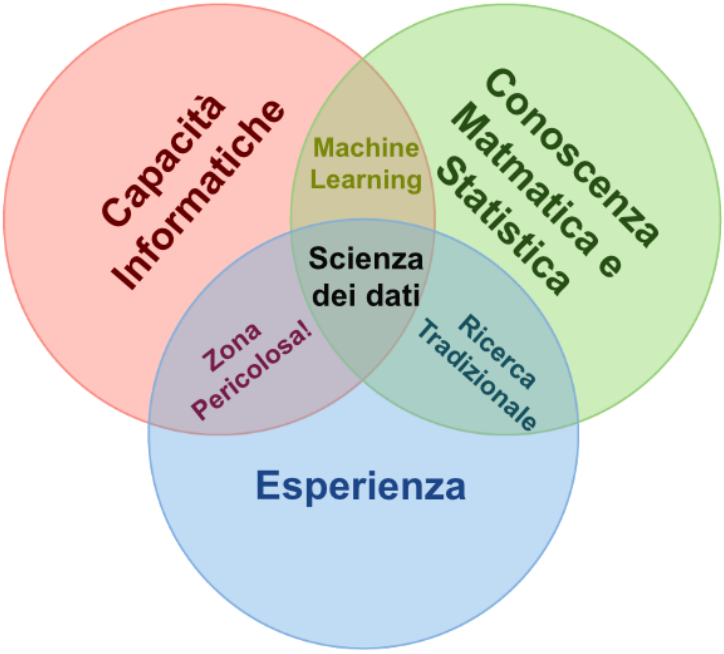


Figura 1.3: Il diagramma di Venn della Data Science ideato da Drew Conway

1.3.1 Ciclo di vita della Data Science: modello CRISP-DM

Il modello CRISP-DM (Figura 1.4) è lo standard metodologico più utilizzato dalla Data Science. Esso fornisce un ciclo di vita strutturato per sviluppare progetti di machine learning e analisi dei dati, trasformando problemi di business in soluzioni concrete. Il modello CRISP-DM è flessibile e può essere facilmente personalizzato. Esso è costituito da sei fasi, ma la loro sequenza non è rigorosa; nella maggior parte dei progetti, infatti, ci si sposta avanti e indietro tra le fasi in base alle necessità. Queste sei fasi sono:

- Business Understanding
    - *Definire il problema:* comprendere qual è l'obiettivo che l'azienda desidera raggiungere.
    - *Tradurre gli obiettivi aziendali in obiettivi tecnici:* questo implica trasformare una necessità aziendale in un problema di machine learning, come classificazione, regressione o clustering.
    - *Stabilire i criteri di successo:* decidere quali metriche o indicatori misureranno il successo del modello.
- Data Understanding

- *Raccolta dei dati:* ottenere le fonti di dati necessarie.
- *Esplorazione iniziale dei dati:* utilizzare analisi statistiche e visualizzazioni per comprendere distribuzioni, valori anomali e schemi.
- *Identificazione dei problemi di qualità:* rilevare dati incompleti, incoerenti o ridondanti che potrebbero influire sulle prestazioni del modello.

▶ Data Preparation

- *Pulizia dei dati:* eliminare duplicati, imputare valori mancanti e correggere errori.
- *Creazione di nuove variabili:* generare attributi derivati che possano essere utili al modello.
- *Normalizzazione e scaling:* adattare le variabili per garantire che siano su intervalli comparabili.
- *Divisione del dataset:* separare i dati in set di addestramento, validazione e test.

▶ Modeling

- *Selezione degli algoritmi:* scegliere i metodi più adatti al problema, come alberi decisionali, reti neurali o metodi ensemble.
- *Ottimizzazione degli iperparametri:* regolare configurazioni, come la profondità degli alberi, i tassi di apprendimento o il numero di epoche.
- *Valutazione iniziale:* utilizzare metriche come precisione, recall, F1 o errore quadratico medio (MSE) per valutare il modello.

▶ Evaluation

- *Revisione delle metriche chiave:* verificare che le prestazioni del modello siano sufficienti rispetto ai criteri di successo.
- *Validazione con dati reali:* testare il modello con dati non utilizzati nell'addestramento.
- *Garanzia di interpretabilità:* valutare se i risultati sono comprensibili e utilizzabili per le parti interessate.
- *Decisione sul passo successivo:* se i criteri di successo non sono soddisfatti, il processo prevede di tornare alla fase di Business Understanding per ridefinire gli obiettivi o l'approccio prima di procedere.

▶ Deployment

- *Integrazione nei sistemi esistenti:* implementare il modello in applicazioni, dashboard o processi.
- *Automazione:* configurare pipeline di dati per aggiornare periodicamente il modello.
- *Monitoraggio e manutenzione:* stabilire sistemi per supervisionare le prestazioni del modello e aggiornarlo quando necessario.



Figura 1.4: Modello CRISP-DM

1.4 Data Analytics

La Data Analytics è il processo con cui si ricavano informazioni da dati che vengono estratti, trasformati e centralizzati per scoprire e analizzare schemi nascosti, relazioni, tendenze, correlazioni e anomalie, oppure per convalidare una teoria o un'ipotesi.

In passato i dati venivano analizzati per prendere decisioni future. Oggi i dati possono essere analizzati per prendere decisioni in tempo reale, individuare trend emergenti e svelare condizioni che non sarebbero evidenti utilizzando processi di gestione dei dati obsoleti.

Implementare la Data Analytics significa dotarsi di un sistema di supporto alle decisioni che permette di agire con maggiore consapevolezza e reattività. Il valore generato si riflette su tutta l'organizzazione, dalla riduzione dei costi operativi al miglioramento della relazione con i clienti. I benefici concreti della Data Analytics si manifestano principalmente nell'ottimizzazione dei processi e delle risorse, permettendo di identificare le aree di miglioramento per aumentare la produttività complessiva. Inoltre, la Data Analytics si rivela fondamentale per ridurre i costi operativi eliminando sprechi e inefficienze, offrendo allo stesso tempo una migliore comprensione del mercato attraverso strumenti avanzati che consentono di analizzare il comportamento dei clienti e adattare le strategie commerciali.

Per sfruttare la Data Analytics, è necessario disporre di strumenti adeguati, capaci di raccogliere, elaborare e rendere fruibili le informazioni. Questi sono:

- *data warehouse*: raccoglie e organizza i dati provenienti da fonti aziendali diverse, rendendoli facilmente accessibili per l'analisi;
- *ETL (Extract, Transform, Load)*: permette di estrarre, trasformare e caricare i dati in un formato omogeneo e utilizzabile;
- *report e dashboard interattive*: consentono di visualizzare i dati in tempo reale, facilitando l'interpretazione delle informazioni;

- *advanced analytics*: tecnologie avanzate come il machine learning permettono di prevedere scenari futuri e ottimizzare le strategie aziendali.

Ogni area aziendale può sfruttare l'analisi dei dati per migliorare efficienza e risultati. A livello generale, l'analisi dei dati si rivela fondamentale per il monitoraggio delle performance e come supporto alle decisioni strategiche, mentre in ambito finanziario guida la pianificazione, il budgeting e la previsione dei flussi di cassa sulla base di dati reali. Parallelamente, le funzioni di marketing e vendite beneficiano dell'analisi di mercato della segmentazione dei clienti e della personalizzazione delle campagne; la produzione ne trae vantaggio per l'ottimizzazione dei processi e la gestione delle supply chain; infine, il settore delle risorse umane può sfruttarla per la gestione del personale, l'analisi delle performance e l'identificazione dei talenti.

1.4.1 Categorie di Data Analytics

All'interno della Data Analytics ci sono quattro categorie generali di analisi (Figura 1.5) che si distinguono sulla base dei risultati che essi producono:

- *Analisi descrittiva*: si concentra sulla descrizione dei dati passati per renderli "digeribili" e offrire un modo per capire e interpretare lo stato del business. Risponde alla domanda generica "che cosa è successo" attraverso la predisposizione di tabelle e grafici, arricchiti dall'uso di statistiche riassuntive (come media, mediana e varianza), semplici trasformazioni e aggregazioni di valori (come indici, conteggi e somme). Il prodotto principale (e anche il più elementare) associato all'analisi descrittiva è il report statico: questo può essere costituito da un semplice file in qualsiasi formato (documenti PDF e fogli Excel sono i più ricorrenti) che viene distribuito su base regolare via e-mail o pubblicato su una piattaforma di distribuzione.

- *Analisi diagnostica*: è un arricchimento dei report descrittivi con indicatori che cercano di spiegare le cause di quello che si osserva nei dati. Utilizzando metodi semplici, come l'analisi di correlazione, le carte di controllo (o control chart) e i test statistici, l'analisi diagnostica può evidenziare degli schemi rilevanti, facendo luce sulle ragioni per cui, per esempio, il business stia andando in un certo modo. Il livello successivo di raffinatezza è costituito dai business alert: in questo caso, i controlli diagnostici vengono eseguiti automaticamente e gli utenti ricevono una notifica solo quando si verifica qualche situazione degna di nota per il business (come, per esempio, il fatto che la quota di mercato di un marchio scenda sotto una certa soglia). In maniera simile, gli strumenti di anomaly detection (rilevazione delle anomalie) sono in grado di ispezionare continuamente i dati alla ricerca di incongruenze e fenomeni atipici che, una volta segnalati, possono essere gestiti opportunamente.

- *Analisi predittiva*: si concentra sulla risposta alle domande supplementari che ci si pone naturalmente dopo aver appreso ciò che è successo in passato, ovvero: "che cosa succederà?". Queste metodologie fanno leva su tecniche analitiche più sofisticate, talvolta basate sull'Intelligenza Artificiale, in quanto dovranno necessariamente andare oltre la semplice descrizione dei fatti storici e avventurarsi nella creazione di modelli probabilistici. Grazie alla predictive analytics si può tentare di svelare le relazioni che intercorrono tra dati diversi ed estrapolarle opportunamente in modo da anticipare un futuro possibile.

- *Analisi prescrittiva*: ci indica direttamente quali azioni mettere in campo per raggiungere un determinato obiettivo, rispondendo, quindi, alla domanda che ogni manager

aziendale si pone continuamente: “come possiamo realizzarlo?”. Per esempio, uno strumento prescrittivo potrebbe simulare un gran numero di scenari alternativi e vagliarli sistematicamente per restituire la "migliore soluzione da seguire per massimizzare il profitto o minimizzare i costi".

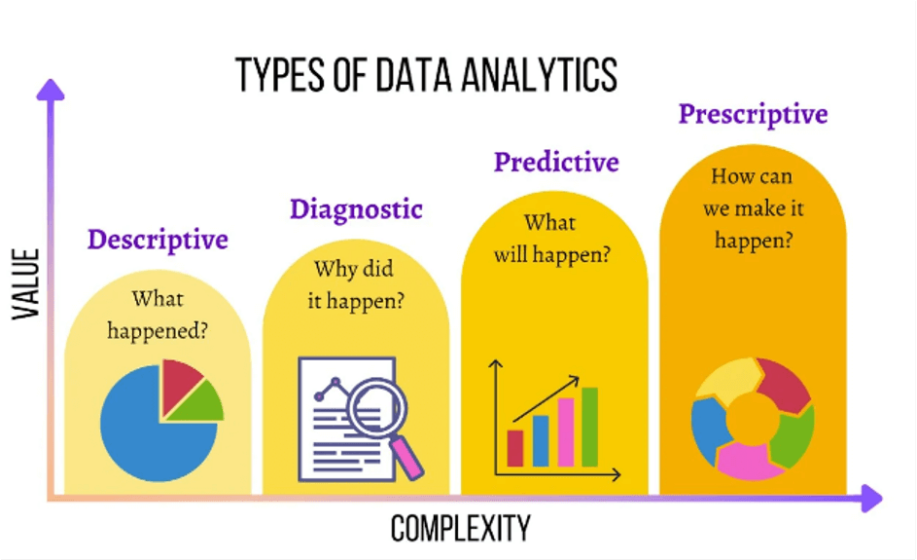


Figura 1.5: Categorie di Data Analytics

1.4.2 Differenze con la Data Analysis

Possiamo definire la Data Analysis come un processo interno alla Data Analytics in cui si ispezionano, puliscono, trasformano e modellano i dati, mentre la Data Analytics utilizza gli insight derivanti da questa analisi per prendere decisioni aziendali migliori. La differenza principale riguarda il livello di astrazione. La Data Analysis è la parte più operativa e tecnica: il suo obiettivo è pulire i dataset grezzi, eseguire query e fare manipolazioni matematiche per estrarre informazioni utili dal rumore di fondo. La Data Analytics, al contrario, rappresenta il quadro strategico più ampio. Il suo compito è prendere quelle stesse informazioni, contestualizzarle all’interno delle dinamiche aziendali e trasformarle in decisioni concrete.

Questi processi non possono fare a meno l’uno dell’altro, poiché analizzare i dati senza uno scopo è, di fatto, privo di senso; d’altra parte prendere decisioni ponderate nella Data Analytics è impossibile senza una fase preliminare di analisi; di conseguenza, l’utilizzo di entrambe si rivela strettamente necessario.

1.4.3 Ciclo di vita della Big Data Analytics

L’analisi dei Big Data si differenzia dall’analisi dei dati tradizionale principalmente per il volume, la velocità e la varietà dei dati elaborati. Per soddisfare i requisiti specifici dell’analisi dei Big Data, è necessaria una metodologia graduale per organizzare le attività e i compiti relativi all’acquisizione, all’elaborazione, all’analisi e al riutilizzo dei dati. Il ciclo di vita della Big Data Analytics è composto da nove fasi, ovvero:

- *Business Case Evaluation*: ogni ciclo di vita dell’analisi dei Big Data deve iniziare con un Business Case ben definito che presenti una chiara comprensione della giustificazione, della motivazione e degli obiettivi dell’analisi stessa. La fase di valutazione del business

case richiede che questo venga creato, valutato e approvato prima di procedere con le attività di analisi vere e proprie.

Questa fase aiuta i responsabili delle decisioni a comprendere le risorse che dovranno essere utilizzate e le sfide aziendali che l'analisi affronterà. L'ulteriore identificazione dei KPI in questa fase può aiutare a definire i criteri e le linee guida per la valutazione dei risultati analitici. Se i KPI non sono immediatamente disponibili, è opportuno impegnarsi a rendere gli obiettivi del progetto SMART, ovvero specifici, misurabili, raggiungibili, rilevanti e con scadenza definita.

Un altro risultato di questa fase è la determinazione del budget necessario per realizzare il progetto di analisi.

- *Data Identification*: la fase di identificazione dei dati è dedicata all'individuazione dei dataset necessari per il progetto di analisi e delle relative fonti. L'identificazione di una più ampia varietà di fonti di dati può aumentare la probabilità di trovare modelli e correlazioni nascosti. Ad esempio, per ottenere informazioni più approfondite, può essere utile identificare il maggior numero possibile di tipologie di fonti di dati correlate, soprattutto quando non è chiaro esattamente cosa cercare.

  A seconda della portata del progetto di analisi e della natura dei problemi aziendali da affrontare, i set di dati necessari e le relative fonti possono essere interni e/o esterni all'azienda.

  Nel caso di dataset interni, in genere viene compilato un elenco di dataset disponibili provenienti da fonti interne, come data mart e sistemi operativi; questo viene poi confrontato con una specifica di dataset predefinita.

  Nel caso di dataset esterni, viene compilato un elenco di possibili fornitori di dati di terze parti, come mercati di dati e dataset pubblicamente disponibili. Alcune forme di dati esterni possono essere incorporate in blog o altri tipi di siti web basati su contenuti, nel qual caso potrebbe essere necessario raccoglierli tramite strumenti automatizzati.

- *Data Acquisition and Filtering*: durante la fase di acquisizione e filtraggio dei dati, questi vengono raccolti da tutte le fonti identificate nella fase precedente. I dati acquisiti vengono, quindi, sottoposti a un filtraggio automatico per la rimozione di quelli corrotti o ritenuti irrilevanti ai fini dell'analisi. I dati classificati come "corrotti" possono includere record con valori mancanti o privi di senso, oppure tipi di dati non validi.

  Sia i dati interni che quelli esterni devono essere memorizzati su disco non appena vengono generati o entrano nel perimetro aziendale. Nel caso dell'analisi batch, questi dati vengono memorizzati su disco prima dell'analisi. Nell'analisi in tempo reale, i dati vengono prima analizzati e poi memorizzati su disco.

  I metadati possono essere aggiunti automaticamente ai dati provenienti da fonti interne ed esterne per migliorarne la classificazione e l'interrogazione. Esempi di metadati aggiunti includono la dimensione e la struttura del dataset, le informazioni sulla fonte, la data e l'ora di creazione o raccolta e le informazioni specifiche sulla lingua. È fondamentale che i metadati siano leggibili automaticamente e trasmessi alle fasi di analisi successive per mantenere la tracciabilità dei dati durante l'intero ciclo di vita, aiutando a stabilire e preservare la loro accuratezza e qualità.

- *Data Extraction*: alcuni dei dati identificati come input per l'analisi potrebbero arrivare in un formato incompatibile con la soluzione Big Data. La necessità di gestire tipologie di dati eterogenee è più probabile con i dati provenienti da fonti esterne. Questa fase è dedicata all'estrazione di dati eterogenei e alla loro trasformazione in un formato utilizzabile dalla soluzione Big Data sottostante per l'analisi dei dati.

L'entità dell'estrazione e della trasformazione necessarie dipende dai tipi di analisi e dalle funzionalità della soluzione Big Data. Ad esempio, l'estrazione dei campi richiesti da dati testuali delimitati, come i file di log dei server web, potrebbe non essere necessaria se la soluzione Big Data sottostante è già in grado di elaborare direttamente tali file. Analogamente, l'estrazione di testo per l'analisi testuale, che richiede la scansione di interi documenti, risulta semplificata se la soluzione di Big Data sottostante è in grado di leggere direttamente il documento nel suo formato nativo.

- *Data Validation and Cleansing*: i dati non validi possono distorcere e falsificare i risultati delle analisi. A differenza dei dati aziendali tradizionali, la cui struttura è predefinita e sono pre-validati, i dati inseriti nelle analisi dei Big Data possono essere non strutturati e privi di qualsiasi indicazione di validità. La loro complessità può, inoltre, rendere difficile definire un insieme di vincoli di validazione adeguati. La fase di convalida e pulizia dei dati è dedicata alla definizione di regole di convalida, spesso complesse, e alla rimozione di eventuali dati non validi noti. Le soluzioni Big Data spesso ricevono dati ridondanti provenienti da diversi dataset; questa ridondanza può essere sfruttata per esplorare dataset interconnessi al fine di assemblare parametri di validazione e colmare le lacune relative ai dati validi mancanti.

  Per l'analisi batch, la convalida e la pulizia dei dati possono essere effettuate tramite un'operazione di ETL offline. Per l'analisi in tempo reale, è necessario un sistema in memoria più complesso per convalidare e pulire i dati man mano che arrivano dalla sorgente.

- *Data Aggregation and Representation*: questa fase è dedicata all'integrazione di più set di dati per ottenere una visione unificata.

  Gli ingenti volumi di dati elaborati dalle soluzioni Big Data possono rendere l'aggregazione un'operazione dispendiosa in termini di tempo e risorse.

  In questa fase è necessario considerare i requisiti futuri per l'analisi, al fine di promuovere il riutilizzo dei dati. Indipendentemente dal fatto che sia necessaria o meno l'aggregazione, è importante comprendere che gli stessi dati possono essere memorizzati in molti formati diversi. Un formato può essere più adatto di un altro per un particolare tipo di analisi.

- *Data Analysis*: questa fase è dedicata all'esecuzione dell'analisi vera e propria, per la quale, in genere, vengono applicate una o più tecniche. Questa fase può essere iterativa, soprattutto se l'analisi è di tipo esplorativo; in tal caso questa viene ripetuta fino a quando non viene individuato il modello o la correlazione appropriata.

  A seconda del tipo di risultato analitico richiesto, questa fase può essere semplice, come interrogare un dataset per calcolare un'aggregazione a scopo di confronto, oppure complessa, come combinare tecniche di data mining e analisi statistiche complesse per scoprire modelli e anomalie o per generare un modello statistico o matematico che rappresenti le relazioni tra le variabili.

  L'analisi confermativa dei dati è un approccio deduttivo in cui la causa del fenomeno oggetto di studio viene proposta a priori (ipotesi). I dati vengono, quindi, analizzati per confermare o smentire l'ipotesi e fornire risposte definitive a domande specifiche. In genere si utilizzano tecniche di campionamento dei dati. Risultati inattesi o anomalie vengono solitamente ignorati, poiché è stata ipotizzata una causa predeterminata.

  L'analisi esplorativa dei dati è un approccio induttivo strettamente correlato al data mining. Non vengono formulate ipotesi o presupposti predeterminati. Al contrario, i dati vengono esplorati attraverso l'analisi per sviluppare una comprensione della causa

del fenomeno. Sebbene possa non fornire risposte definitive, questo metodo offre una direzione generale che può facilitare la scoperta di modelli o anomalie.

- *Data Visualization*: la capacità di analizzare enormi quantità di dati e ricavarne informazioni utili ha poco valore se gli unici in grado di interpretarne i risultati sono gli analisti.

  La fase di visualizzazione dei dati è dedicata all'utilizzo di tecniche e strumenti di visualizzazione per comunicare graficamente i risultati dell'analisi, consentendo un'interpretazione efficace da parte degli utenti aziendali.

  I risultati ottenuti completando questa fase offrono agli utenti la possibilità di eseguire analisi visive, consentendo di scoprire risposte a domande che non hanno ancora nemmeno formulato. Gli stessi risultati possono essere presentati in diversi modi, il che può influenzarne l'interpretazione.

- *Utilization of Analysis Results*: una volta che i risultati dell'analisi sono stati resi disponibili agli utenti per supportare il processo decisionale, ad esempio tramite dashboard, potrebbero presentarsi ulteriori opportunità per utilizzare tali risultati. Questa fase è dedicata a determinare come e dove i dati di analisi elaborati possono essere ulteriormente sfruttati.

  A seconda della natura dei problemi di analisi affrontati, i risultati possono produrre "modelli" che racchiudono nuove intuizioni e comprensioni sulla natura dei modelli e delle relazioni presenti nei dati analizzati. Un modello può assumere la forma di un'equazione matematica o di un insieme di regole. I modelli possono essere utilizzati per migliorare la logica dei processi aziendali e la logica dei sistemi applicativi.

Nella Figura 1.6 viene fornita una rappresentazione grafica delle varie fasi che compongono il ciclo di vita della Big Data Analytics.

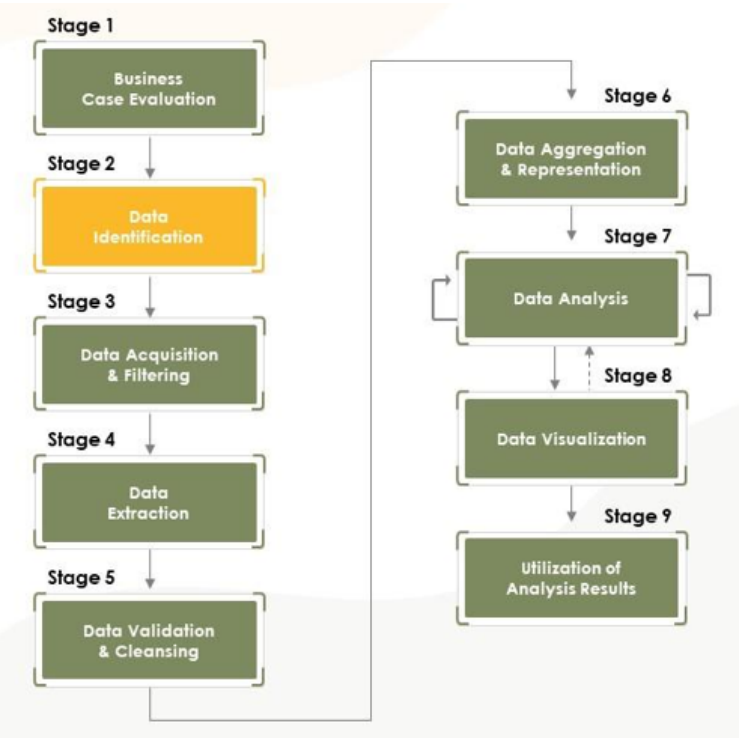


Figura 1.6: Ciclo di vita Big Data Analytics

1.5 Confronto tra Data Science e Data Analytics

Sebbene la Data Science e la Data Analytics possano sembrare molto simili tra loro, in realtà la seconda si concentra sull'interpretazione dei dati esistenti per identificare i trend e generare insight utili per i decision maker aziendali, mentre la prima comprende il machine learning, la creazione di modelli e sistemi automatizzati che prevedono i risultati futuri.

Nella Tabella 1.1 vengono messe a confronto queste due discipline sotto alcuni aspetti.

Dimensione	**Data Analytics**	**Data Science**
Domanda chiave	Che cosa è successo?	Che cosa succederà?
Tipo di dati principale	Dati strutturati	Dati strutturati e non strutturati
Output principale	Dashboard, report, insight di valore	Modelli predittivi, algoritmi
Strumenti chiave	SQL, Excel, Tableau, Power BI	Python, R, Spark, MLflow
Percorso di studi	Analytics, statistica, business	Informatica, matematica
Ruoli tipici	Analytics Engineer, BI Analyst	Data Scientist, ML Engineer

Tabella 1.1: Confronto tra Data Science e Data Analytics

I Data Analyst trascorrono le loro giornate a interrogare dati rilevanti, eseguire analisi descrittive, creare dashboard Power BI o Tableau e aiutare i team a interpretare i dati per individuare pattern. I deliverable dell'analisi dei dati (come sintesi dei trend, output di data visualization e raccomandazioni strutturate) aiutano i decision maker aziendali a prendere rapidamente decisioni basate sui dati. Tra i ruoli più comuni nella Data Analytics figurano Analytics Engineer, Reporting Analyst e BI Analyst.

I Data Scientist progettano esperimenti, sviluppano algoritmi, applicano tecniche di machine learning e creano modelli predittivi per risolvere problemi complessi su larga scala. I loro deliverable (come modelli di churn in produzione, motori di raccomandazione e previsioni della domanda) automatizzano direttamente le decisioni. Tra i ruoli più comuni figurano Applied Scientist, ML Engineer o Research Scientist. Entrambe le discipline richiedono solide competenze tecniche, anche se i data scientist necessitano di una maggiore profondità tecnica e di competenze di modellazione più avanzate.

Le funzioni di Data Analytics e Data Science si integrano in diversi punti. I Data Analyst definiscono il problema aziendale e stabiliscono le metriche di riferimento; i data scientist creano quindi modelli predittivi utilizzando i dati puliti e ben organizzati che i team di analytics hanno già validato. Nelle piccole aziende, un singolo analista con competenze intermedie ed esperienza nella modellazione dei dati può coprire entrambe le funzioni. I team più grandi beneficiano di protocolli di handoff documentati, in modo che ciascun gruppo possa lavorare in modo iterativo senza compromettere le attività a valle.

CAPITOLO 2

Introduzione a Power BI

In questo capitolo viene introdotto Power BI, la piattaforma di Business Intelligence di Microsoft. Oltre a evidenziare vantaggi e limiti, verranno analizzate le sue modalità di utilizzo e i suoi componenti principali, sia riguardo all'architettura sia riguardo al flusso operativo per la realizzazione di report e dashboard. Infine, il capitolo si soffermerà sul processo di Data Visualization e sui linguaggi di programmazione integrati nella piattaforma.

2.1 Power BI

Microsoft Power BI è un software di Business Intelligence che consente agli utenti di creare report interattivi in grado di semplificare la lettura e la condivisione dei dati.

Grazie all'utilizzo di strumenti come Microsoft Power Query e all'interazione con Microsoft Fabric, la piattaforma può connettersi a un'ampia gamma di sorgenti di dati differenti e gestire flussi complessi, supportando l'analisi di grandi volumi di informazioni. Power BI offre una suite completa di funzionalità per l'analisi dei dati aziendali, consentendo alle organizzazioni di trasformare informazioni grezze in insight preziosi.

Nel corso degli anni, il software ha aggiunto funzionalità di Intelligenza Artificiale e machine learning. Molte di queste funzionalità, quali il rilevamento delle anomalie o la ricerca dei fattori chiave, sono integrate, mentre altre sfruttano la piattaforma di Machine Learning fornita da Microsoft attraverso Azure Machine Learning. Le funzionalità AI rendono Power BI uno strumento dinamico e flessibile, in grado di coprire necessità diverse, in ambiti industriali, scientifici e amministrativi.

La piattaforma presenta una struttura cloud-oriented, costituita da componenti desktop, cloud services e app mobile.

Il componente cloud, comunemente identificato come Power BI Service (che verrà approfondito nella Sezione 2.2.1), è il cuore del sistema ed è centrale nell'ottica di condivisione dei contenuti tra tutte le persone coinvolte nelle attività di analisi.

2.1.1 Vantaggi e limiti

Power BI emerge come un leader nel contesto della Business Intelligence, superando costantemente concorrenti come Tableau, QlikView e Looker. Una delle sue caratteristiche principali è la perfetta integrazione con l'ecosistema Microsoft, che migliora la collaborazione

e consente agli utenti di sfruttare le competenze esistenti. I principali vantaggi che permettono di distinguere Power BI rispetto agli altri software sono:

- *Possibilità di fornire Business Intelligence a qualsiasi organizzazione*: la piattaforma promuove una cultura orientata ai dati nelle organizzazioni di tutte le dimensioni. Una cultura data-driven dà priorità alle decisioni basate sui dati piuttosto che sull'intuito o sulle sensazioni. Power BI facilita questo compito fornendo l'accesso a report e dashboard in tempo reale, consentendo a tutti i membri di un'organizzazione di prendere decisioni informate basate su dati affidabili.

- *Robuste funzioni di sicurezza*: Power BI è dotato di una suite completa di funzioni di sicurezza che aiutano a proteggere i dati sensibili e a soddisfare i vari requisiti di conformità. Ad esempio, Cloud App Security di Microsoft offre un'ampia visibilità e un controllo completo sul movimento dei dati, utilizzando analisi avanzate per rilevare e combattere le minacce informatiche in tutti i servizi cloud. Inoltre, le etichette di sensibilità di Power BI aiutano gli amministratori a contrassegnare chiaramente i dati sensibili, garantendo che vengano gestiti in modo appropriato. La piattaforma supporta anche la sicurezza a livello di riga (RLS), che consente agli sviluppatori di assegnare specifiche autorizzazioni di accesso ai dati a diversi utenti, garantendo che le informazioni sensibili siano accessibili solo al personale autorizzato.

- *Integrazione nativa dei dati*: Power BI eccelle nella connessione a un'ampia gamma di fonti di dati. Che si tratti di sorgenti come Excel e CSV, di database come Oracle e Snowflake o di dati provenienti dal web come Salesforce e Adobe Analytics, Power BI si integra perfettamente. Particolarmente degna di nota è l'integrazione con Excel, che consente agli utenti di trasferire le query, i modelli di dati e i report di Excel esistenti su Power BI senza sforzo, senza dover imparare una nuova applicazione o un nuovo linguaggio. Ciò è particolarmente vantaggioso per i processi aziendali ad alta intensità di dati, come quelli utilizzati nel marketing.

- *Capacità di Intelligenza Artificiale*: incorpora funzioni avanzate di Intelligenza Artificiale che consentono agli utenti di estrarre più valore dai loro dati. Queste funzionalità comprendono l'elaborazione del linguaggio naturale, gli algoritmi di apprendimento automatico e le analisi predittive, che aiutano gli utenti a prevedere i risultati e a prendere decisioni più informate. Sfruttando questi potenti strumenti, le aziende possono trasformare i dati grezzi in risorse strategiche, promuovendo l'innovazione e mantenendo un vantaggio competitivo sul mercato.

- *Miglioramento continuo*: Power BI viene regolarmente aggiornato con nuove funzionalità e miglioramenti, garantendo agli utenti l'accesso agli strumenti e alle funzionalità più recenti. Questa dedizione allo sviluppo continuo fa sì che Power BI non solo tenga il passo con l'evoluzione delle esigenze delle aziende, ma anticipi anche le tendenze future dell'analisi dei dati. Gli aggiornamenti regolari includono miglioramenti delle prestazioni, nuovi connettori di dati, opzioni di visualizzazione migliorate e funzionalità avanzate di Intelligenza Artificiale, il tutto progettato per fornire un'esperienza utente più potente e intuitiva.

- *Visualizzazioni facili da usare*: Power BI rende la visione dei dati intuitiva e accessibile. Gli utenti possono creare un'ampia gamma di visualizzazioni, le quali aiutano a trasformare i dati grezzi in storie avvincenti che favoriscono un migliore processo decisionale. La natura interattiva di queste visualizzazioni consente agli utenti di applicare filtri e vedere i dati aggiornati in tempo reale, fornendo approfondimenti e favorendo una migliore comprensione dei dati.

- *Condivisione efficiente dei contenuti con le app*: le app sono un modo conveniente per gli sviluppatori di report di distribuire i contenuti agli utenti, sia che si tratti di un grande gruppo, di un'intera organizzazione o di una singola persona. In questo modo è più facile garantire che le persone giuste abbiano accesso alle informazioni giuste quando ne hanno bisogno.

Nonostante i numerosi vantaggi, Power BI presenta alcune limitazioni che ne possono condizionare l'utilizzo. Tra queste citiamo:

- *costi nascosti*: sebbene l'accesso iniziale sembri economico, molte funzionalità avanzate richiedono aggiornamenti costosi; inoltre, con l'introduzione di Microsoft Fabric, i costi complessivi (TCO) sono aumentati in modo significativo;
- *gestione complessa*: l'uso di strumenti desktop, necessario per creare report, può portare a una governance inefficiente, a una diffusione di file ridondanti tra gli utenti e a una duplicazione dei flussi di dati;
- *dipendenza dal cloud*: Power BI offre opzioni limitate per implementazioni on-premise o multi-cloud, favorendo un lock-in con i servizi cloud di Azure;
- *modellazione limitata*: con l'aumento delle dimensioni e della complessità dei dati, Power BI può richiedere delle query live per mantenere la reattività, compromettendo le prestazioni.

2.1.2 Modalità di utilizzo

L'adozione di Power BI all'interno del contesto aziendale risponde a esigenze diverse in base alla professione coinvolta. Di seguito vengono analizzati i principali ambiti di utilizzo.

Data Analyst

Per gli analisti dei dati, Power BI rappresenta uno degli strumenti principali, soprattutto per la Data Analysis, poiché agevola la raccolta, l'analisi e la comunicazione efficace dei dati. Questo aspetto è fondamentale, in quanto il lavoro dell'analista non si limita all'analisi, ma include anche la capacità di raccontare i dati. Power BI consente di trasformare numeri complessi in storie visive, rendendo le informazioni accessibili anche a chi non ha competenze tecniche. Inoltre, esso permette di connettere fonti diverse, creare grafici interattivi e far emergere insight nascosti, fino a realizzare report e dashboard aziendali capaci di mantenere aggiornato l'intero team. L'ambiente Power Query nella versione per Desktop fornisce, inoltre, degli strumenti per organizzare in tempi molto rapidi e a bassi costi un'ingente mole di dati.

Reparto IT

Power BI ottimizza la gestione del lavoro di una singola persona, ma aiuta ancora di più le grandi aziende che devono tenere sotto controllo numerosi reparti, ognuno con molti flussi di dati. Un primo vantaggio strategico risiede nel fatto che spesso le grandi realtà lavorano già con i prodotti Microsoft; questo offre una perfetta integrazione tra il background già esistente e un sistema innovativo che può portare un netto miglioramento nella qualità del lavoro. Tutti i dati e i report possono essere gestiti dal reparto IT, supportato dalla sicurezza garantita da Power BI. La centralizzazione dei dati su Power BI permette al reparto IT di mantenere una supervisione costante sulle dinamiche interne dell'azienda, prevenendo il mancato aggiornamento delle informazioni e garantendo il monitoraggio di eventuali anomalie.

Sviluppatori

Gli sviluppatori possono incorporare Power BI in siti web, applicazioni e portali, creando delle dashboard interattive che offrono un miglioramento dell'esperienza utente. Oltre a ciò, essi hanno l'opportunità di creare oggetti visivi interattivi e visualizzazioni di dati personalizzate grazie a piani di lavoro, aggiornamento dei dati e diverse funzionalità che possono essere automatizzate per facilitare il lavoro degli esperti.

2.2 Componenti

2.2.1 Architettura

L'ecosistema di Power BI è costituito da componenti sia locali sia cloud che collaborano tra loro per offrire funzionalità avanzate per quasi tutte le esigenze di analisi e visualizzazione dei dati. Nella Figura 2.1 viene rappresentata l'architettura di Power BI. I suoi principali componenti sono:

- *Power BI Desktop*;
- *Power BI Service*;
- *Power BI Gateway*;
- *Power BI Mobile Apps*;
- *Power BI Visuals Marketplace*.

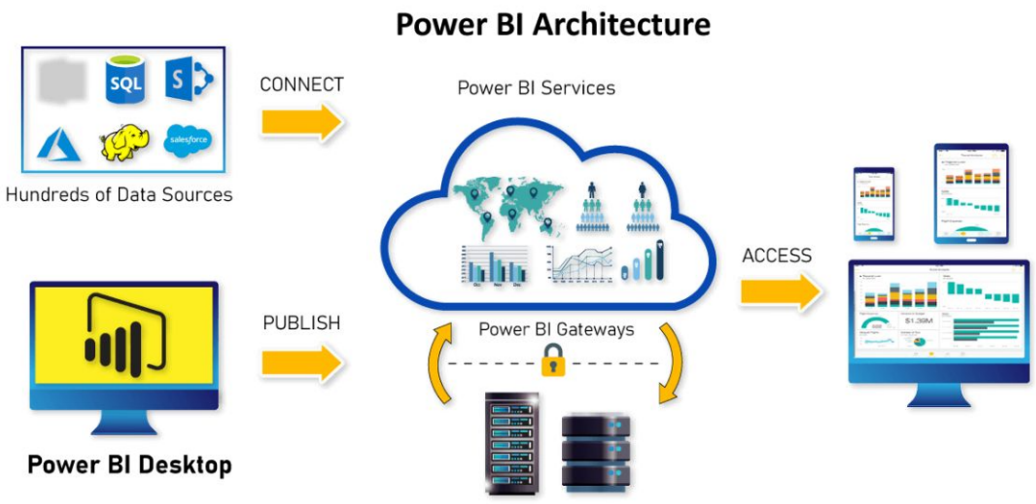


Figura 2.1: Architettura di Power BI

Power BI Desktop

Power BI Desktop è la versione desktop della piattaforma Power BI, disponibile solo per computer Windows. Se si utilizza Mac o Linux è possibile adoperarla tramite una macchina virtuale o un desktop remoto. Essa è gratuita, autonoma e offre funzionalità di trasformazione, modellazione, visualizzazione e creazione di report sui dati; inoltre, Power BI Desktop

consente un'esperienza di Business Intelligence self-service tramite un'interfaccia intuitiva. Con Power BI Desktop, gli utenti possono:

- connettersi alle sorgenti dati e importare questi ultimi;
- trasformare, pulire e ristrutturare i dati secondo le loro necessità;
- modellare e analizzare i dati creando relazioni e calcoli;
- progettare report interattivi con varie visualizzazioni, come grafici, mappe e widget;
- pubblicare report e set di dati in Power BI Service per condividerli con altri.

Power BI Service

Power BI Service è una piattaforma SaaS per collaborare, pubblicare, gestire e distribuire report e dashboard. I contenuti creati dagli utenti tramite Power BI Desktop possono essere caricati qui. Le principali funzionalità offerte da Power BI Service includono:

- condivisione di dashboard e report interattivi con altri membri di un'organizzazione;
- discussione sui dati tramite commenti e annotazioni;
- pubblicazione di set di dati e report personalizzati per l'intera organizzazione;
- aggiornamento automatico dei report con i dati più recenti;
- controllo degli accessi e delle autorizzazioni in conformità alle politiche di sicurezza;
- monitoraggio delle metriche di utilizzo e dei registri di controllo;
- integrazione di contenuti in app e applicazioni di terze parti.

Power BI Service, quindi, è l'hub centrale in cui tutti gli utenti finali possono accedere e interagire con i contenuti di Power BI tramite un browser web o un'app mobile.

Power BI Gateway

Power BI Gateway rappresenta un gateway dei dati locale e opzionale, necessario se la sorgente dei dati si trova all'interno di una rete privata e deve essere mantenuta completamente al sicuro. Il gateway funge da ponte per consentire la comunicazione tra le sorgenti dei dati locali e il servizio Power BI nel cloud. Esso fornisce un accesso controllato ai dati senza esporli su Internet. Microsoft fornisce due tipi di gateway:

- *Personal gateway*: esso consente agli utenti di connettersi alle sorgenti dei dati nel proprio ambiente locale per il loro aggiornamento e per realizzare le DirectQuery; esso viene installato sul computer locale dell'utente.
- *Enterprise gateway*: esso consente la connettività a livello di organizzazione verso le sorgenti dei dati locali e offre funzionalità di condivisione; esso viene installato all'interno del firewall aziendale.

Power BI Mobile Apps

Per consentire la fruizione degli strumenti di analisi anche in mobilità, Power BI offre app native per iOS, Android e Windows. Queste app permettono un facile accesso a dashboard e report di Power BI da dispositivi mobili. Le funzionalità principali includono:

- dashboard ottimizzate per schermi più piccoli;
- notifiche push sugli aggiornamenti dei dati;
- accesso offline tramite cache.

Power BI Visuals Marketplace

Power BI Visuals Marketplace (noto ufficialmente come Microsoft AppSource) è lo store integrato di Microsoft in cui si possono scaricare e aggiungere oggetti visivi personalizzati per la realizzazione di report tramite grafici avanzati non presenti di default in Power BI. In questo store sono presenti sia elementi visivi gratuiti sia soluzioni a pagamento tramite licenze. La piattaforma offre, inoltre, la possibilità di filtrare gli oggetti visivi per identificare quelli certificati da Microsoft, i quali garantiscono il superamento di rigidi test di sicurezza, assicurano che non vi sia invio di dati all'esterno dell'ambiente aziendale e confermano la piena compatibilità con tutte le funzionalità di esportazione (ad esempio, in PowerPoint o in PDF).

2.2.2 Operatività

L'operatività in Power BI è costituita da un flusso di azioni, le quali possono essere sintetizzate nei seguenti punti:

- connettersi alle sorgenti dei dati;
- eseguire operazioni di Data Cleaning, ovvero trasformare e pulire i dati per creare un modello;
- creare oggetti visivi che forniscono una rappresentazione dei dati;
- creare report che raccolgono gli oggetti visivi in una o più pagine;
- condividere i report con altri utenti tramite il servizio Power BI.

Connettersi ai dati

La piattaforma dispone di connettori nativi per diverse sorgenti di dati, tra cui:

- *file locali*: Excel, Testo/CSV, XML, JSON;
- *database*: SQL Server, Oracle, MySQL, PostgreSQL, MariaDB, MongoDB, ecc.;
- *Microsoft Fabric*: modelli semantici di Power BI, Flussi di dati, Lakehouse, Database KQL;
- *Power Platform*: flussi di dati Power BI (legacy), Common Data Service (legacy), Dataverse;
- *cloud Azure*: database SQL di Azure, Database di Azure Analysis Services, Database di Azure per PostgreSQL, Azure Databricks, ecc.;

- *Online Services*: SharePoint Online, Microsoft Exchange Online, Dynamics 365 Online, GitHub;
- *altre sorgenti*: Web, Microsoft Exchange, Spark, ecc.

L'interfaccia di Power BI rende possibile configurare le sorgenti dei dati tramite l'opzione *Recupera dati* presente nella barra multifunzione *Home* (Figura 2.2).

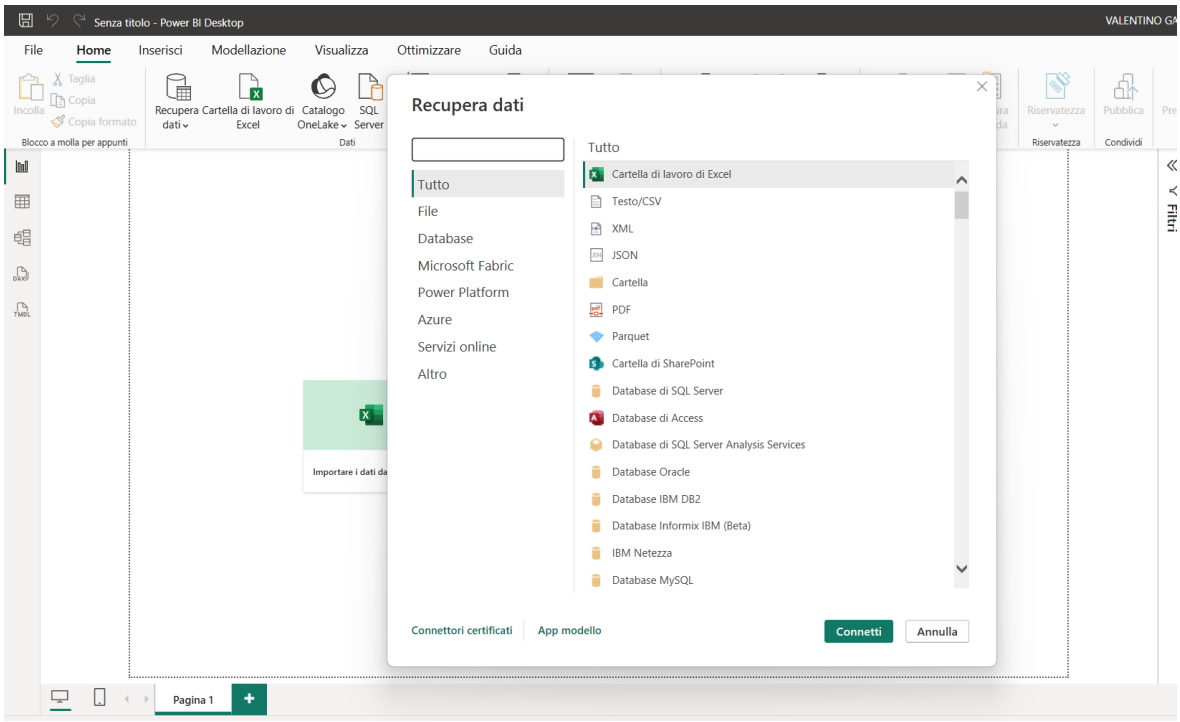


Figura 2.2: Connessione ai dati

Data Cleaning

I dati grezzi potrebbero non essere sempre pronti per un'analisi diretta. Power BI mette a disposizione funzionalità di trasformazione e modellazione dei dati per prepararli.

Una volta connessa una sorgente dei dati, lo strumento Power Query Editor (Figura 2.3) viene utilizzato per eseguire varie operazioni di trasformazione e pulizia dei dati, quali ad esempio:

- rimozione o aggiunta di tabelle;
- unione di tabelle;
- operazione di pivot e unpivot sulle tabelle;
- modifica dei nomi di tabelle e colonne;
- modifica dei tipi di dati;
- rimozione di colonne;
- aggiunta di colonne calcolate o condizionali;
- filtraggio di colonne e righe;

- rimozione di errori e di valori duplicati;
- gestione dei valori nulli.

L'accesso a Power Query Editor avviene tramite il comando *Trasforma dati* presente nella schermata *Home*. Oltre a tenere traccia delle modifiche effettuate sui dati, esso mette a disposizione un *Editor avanzato* per interagire con il linguaggio M generato da Power BI nelle operazioni di Data Cleaning.

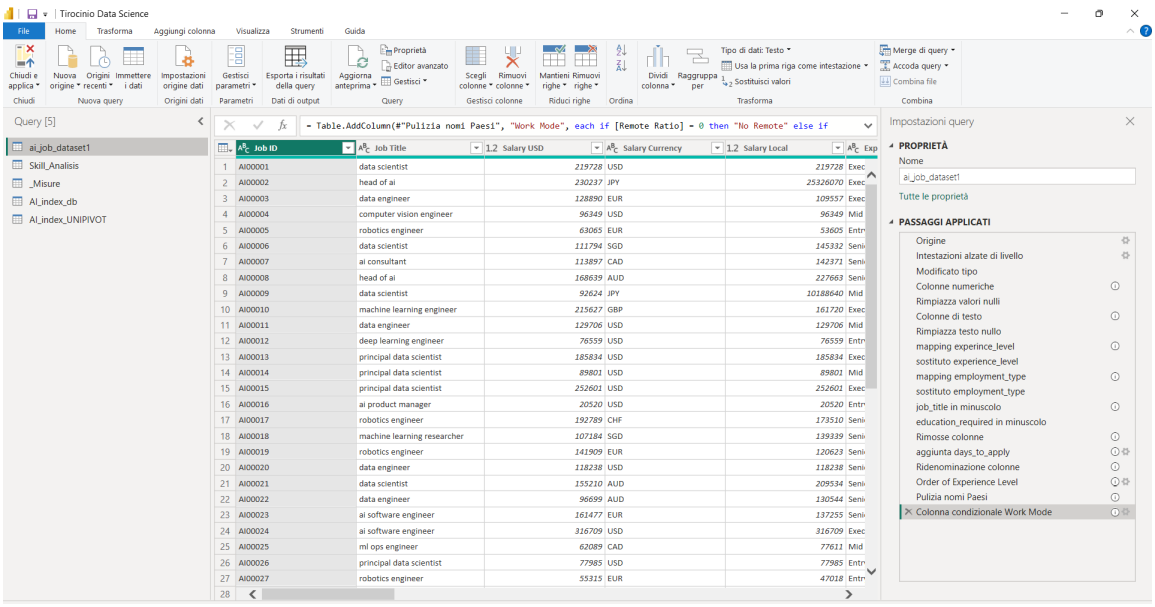


Figura 2.3: Power Query Editor

Creare oggetti visivi

Una volta creato un modello dei dati, si passa alla creazione degli oggetti visivi. Power BI Desktop mette a disposizione un pannello di visualizzazioni (Figura 2.4) per l'inserimento di grafici a linee, a torta, matrici, tabelle, oppure oggetti visivi di terze parti provenienti dall'AppSource di Microsoft; grazie a questi strumenti di visualizzazione, la piattaforma riesce a rappresentare i dati in maniera chiara e comprensibile per gli utenti.

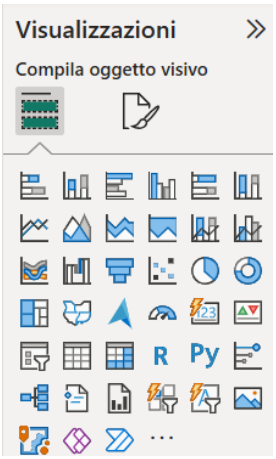


Figura 2.4: Pannello di visualizzazioni

Creare report

I diversi oggetti visivi, creati al passo precedente vengono raccolti e organizzati su una o più pagine per realizzare dei report.

Mentre i report contengono oggetti visivi collegati ad un unico dataset, le dashboard sono costituite da una singola pagina che integra oggetti visivi provenienti da più dataset. Le visualizzazioni mostrate nella dashboard sono chiamate riquadri; questi vengono aggiunti ad una dashboard a partire dai report. Una dashboard può essere considerata una via d'accesso ai report e ai set di dati sottostanti.

Nella Figura 2.5 viene mostrata la relazione tra le dashboard ed i report.

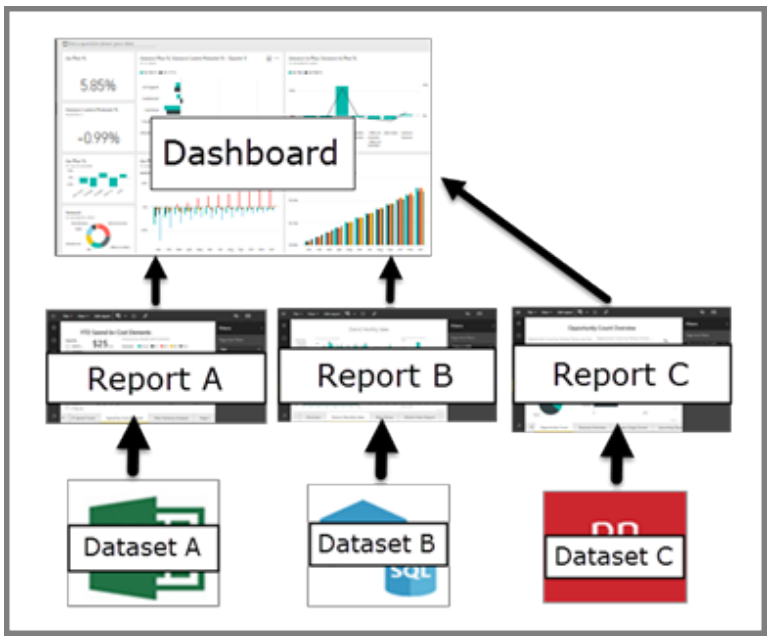


Figura 2.5: Relazione tra dashboard e report

Condivisione dei report

Una volta terminato il report, esso può essere pubblicato su Power BI Service per potervi accedere e condividere online. Questa operazione avviene grazie al comando *Pubblica* nella scheda *Home*, come mostrato nella Figura 2.6.



Figura 2.6: Comando di condivisione dei report

2.3 Data Visualization

La Data Visualization è definita come l'esplorazione visuale/interattiva, e la relativa rappresentazione grafica, di dati di qualunque dimensione (small e big data), natura e origine.

Essa permette, in estrema sintesi, a manager e decision maker di identificare fenomeni e trend che risultano invisibili ad una prima analisi dei dati.

La Data Visualization è nata per esplorare e analizzare i dati in maniera visuale e avere una vista di sintesi sul business che, all'occorrenza, può essere approfondita. Inoltre, essa consente di:

- identificare le aree di business in cui apportare modifiche o miglioramenti;
- comprendere i fattori chiave che influenzano il comportamento dei clienti;
- elaborare un più efficace posizionamento dei prodotti;
- effettuare previsioni sui volumi di vendita.

Power BI Desktop mette a disposizione tre tipi di visualizzazione (Figura 2.7), ognuna delle quali ha uno scopo preciso durante l'analisi:

- *Vista Report*: consente di usare le query create dall'utente per generare visualizzazioni ed organizzarle in una o più pagine.
- *Vista Dati*: consente di visualizzare i dati caricati nel report in un formato di modello di dati, su cui è possibile aggiungere misure, creare nuove colonne e gestire le relazioni.
- *Vista Relazioni*: consente di ottenere una rappresentazione grafica delle relazioni che sono state definite nel modello di dati e di gestirle o modificarle in base alle esigenze.



Figura 2.7: Tre tipi di visualizzazione in Power BI Desktop: Vista Report, Vista Dati, Vista Relazioni

Infine, per rendere interattiva la visualizzazione, Power BI Desktop mette a disposizione degli utenti un'area di gestione dei filtri che consente di limitare e selezionare i dati di visualizzazione, mostrando solo le informazioni rilevanti in base a determinati criteri.

2.3.1 Vista Report

La Vista Report è utilizzata per creare diverse pagine di report con visualizzazioni; questa offre un'esperienza di progettazione simile alla visualizzazione di modifica del report nel

servizio Power BI. Essa permette di usare le query e modellare i dati per assicurarsi che essi supportino le informazioni dettagliate migliori nei report. Quando viene visionato un report, l'interfaccia mostra tre pannelli laterali per la personalizzazione e modifica di quest'ultimo (Figura 2.8). I tre pannelli sono:

- *Pannello dei Filtri*: consente di visualizzare, impostare e modificare filtri persistenti all'interno dei report, agendo a differenti livelli: di pagina, di report, di drill-through e visivo. Tale pannello offre la flessibilità per poter applicare criteri di filtraggio basati su campi che non sono direttamente inclusi o rappresentati negli elementi visivi del report.
- *Pannello delle Visualizzazioni*: consente di creare la visualizzazione tramite la scelta degli oggetti visivi. Quando viene effettuata la scelta della visualizzazione, questa può essere modificata tramite tre schede: Campi, Formato e Analisi. I *Campi* servono per controllare gli input necessari alla visualizzazione, come valori, categorie, legenda e assi. Il *Formato* serve per modificare la grafica della visualizzazione, come colori, titoli, etichette, bordi e sfondo. L'*Analisi* consente di inserire elementi di analisi automatica, come linee di tendenza, di media, intervalli di riferimento, ecc.
- *Pannello dei Campi*: visualizza le tabelle, le cartelle e i campi dei dati disponibili (comprese anche le misure DAX calcolate) per la creazione di visualizzazioni. Da questa scheda i campi vengono selezionati o trascinati per essere utilizzati nel pannello filtri e visualizzazioni.

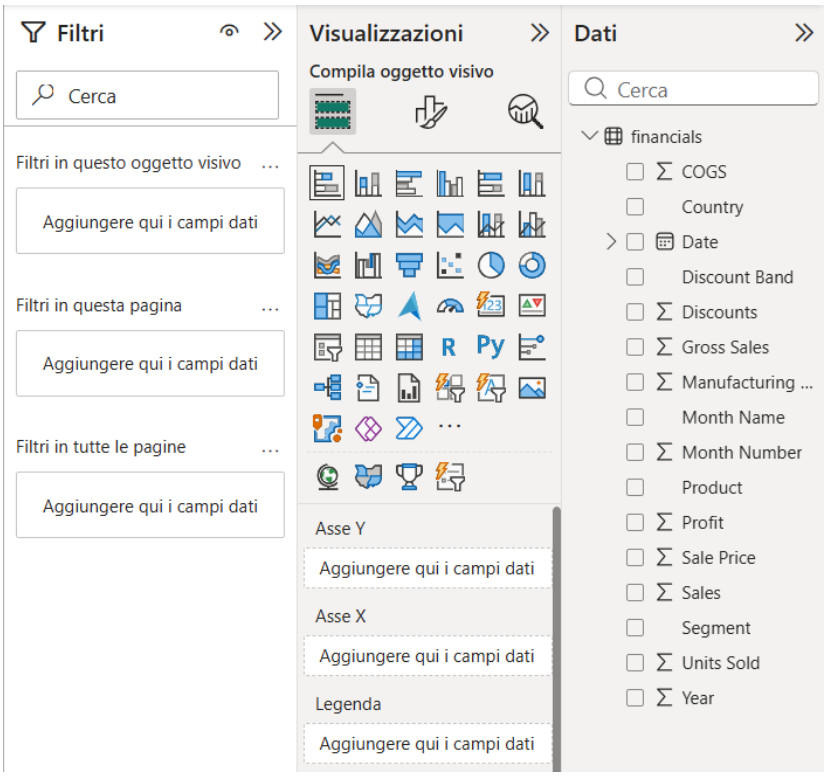


Figura 2.8: Vista Report: pannello Filtri, Visualizzazioni e Campi

2.3.2 Gestione dei Filtri

Power BI offre diversi modi per filtrare i dati nei report. Tutti gli oggetti visivi possono essere filtrati usando il pannello Filtri, che consente di applicare filtri a livello di oggetto visivo, pagina o report.

Per un'interazione più diretta, Power BI mette a disposizione lo slicer, un particolare oggetto visivo per fornire controlli di filtraggio sul frame. Gli slicer sono disponibili in più formati, tra cui pulsante, elenco, elenco a discesa e intervallo di date; questi vengono visualizzati nell'area di disegno del report per facilitare l'accesso.

2.3.3 Vista Dati

La Vista Dati consente di esaminare, esplorare e comprendere i dati nel modello di Power BI Desktop; essa è diversa dal modo in cui si visualizzano tabelle, colonne e dati nell'editor di Power Query. Con la Vista Dati i dati vengono esaminati dopo che sono stati caricati nel modello. Quando si modellano i dati, a volte si vuole vedere cosa sia effettivamente presente in una tabella o colonna senza creare un oggetto visivo nell'area di disegno del report. Questa possibilità è particolarmente utile quando si creano misure e colonne calcolate oppure quando è necessario identificare un tipo o una categoria di dati. Gli elementi fondamentali della vista sono:

- *griglia di dati*, per mostrare la tabella selezionata e tutte le colonne e le righe in essa contenute;
- *barra della formula*, per inserire formule DAX per le misure e le colonne calcolate;
- *ricerca*, per cercare una tabella o una colonna nel modello;
- *elenco campi*, per selezionare una tabella o una colonna da visualizzare nella griglia dei dati.

L'utilità di questa modalità risiede nella possibilità di esaminare accuratamente le misure calcolate a partire da colonne personalizzate, accertandosi che non vi siano valori non validi, nulli o errori. Svolgere questo tipo di controllo preliminare garantisce l'integrità del modello, ponendosi come un passo necessario prima di dedicarsi alla progettazione dell'interfaccia grafica dei report. Nella Figura 2.9 viene mostrato un esempio di Vista Dati.

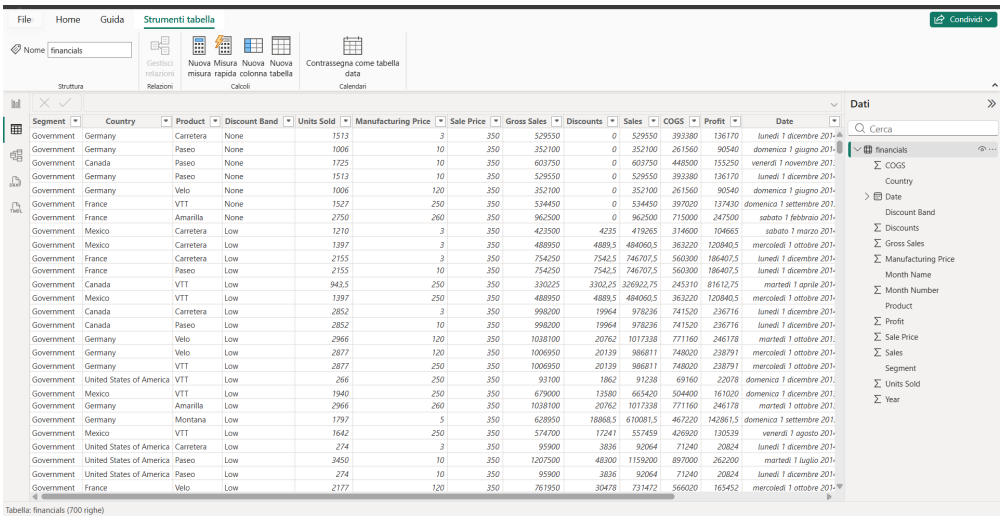


Figura 2.9: Esempio di Vista Dati

2.3.4 Vista Relazioni

Quando si hanno più tabelle, è probabile che si eseguano analisi relative ai dati inclusi in ognuna di esse. Le relazioni tra le tabelle sono necessarie per calcolare con precisione i risultati e visualizzare le informazioni corrette nei report. Nella maggior parte dei casi non è necessario eseguire alcuna operazione; la funzionalità di rilevamento automatico crea automaticamente le varie relazioni. Talvolta, tuttavia, può essere necessario creare manualmente le relazioni, oppure apportarvi alcune modifiche; ciò può essere fatto all'interno della Vista Relazioni. Le operazioni fondamentali che si possono compiere grazie a questa vista sono:

- creare nuove relazioni tra tabelle;
- modificare la cardinalità, come Molti a uno, Uno a uno, Uno a molti e Molti a molti;
- modificare la direzione del filtro incrociato (cross filter direction);
- impostare una relazione come attiva.

Nella Figura 2.10 viene mostrato un esempio di Vista Relazioni.

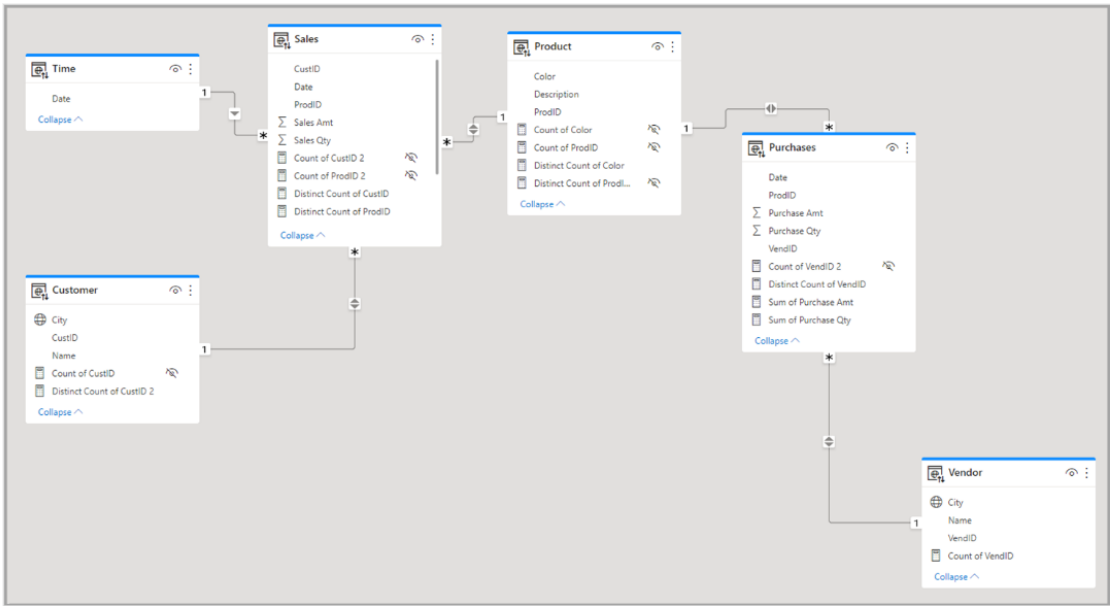


Figura 2.10: Esempio di Vista Relazioni

2.4 Linguaggi di programmazione

Per trasformare i dati grezzi in informazioni strategiche e visualizzazioni d'impatto, Power BI si affida a un ecosistema di linguaggi di programmazione e di scripting. Sebbene la piattaforma offra un'interfaccia grafica altamente intuitiva, l'uso dei linguaggi consente di superarne i limiti, offrendo il pieno controllo sulla manipolazione e sulla modellazione dei dati. Questi linguaggi si dividono tra componenti native del software e integrazioni esterne; i principali sono:

- *DAX*: acronimo di Data Analysis Expressions; è un linguaggio per implementare formule progettato per eseguire calcoli avanzati sui dati in modelli tabulari, come quelli creati in Analysis Services. DAX estende le funzionalità di Excel, consentendo agli utenti di

eseguire calcoli complessi, creare misure, colonne e tabelle personalizzate e analizzare relazioni tra i dati per sviluppare analisi più approfondite. La sua sintassi combina elementi di Excel con l'algebra relazionale, offrendo un linguaggio ricco e intuitivo per la manipolazione dei dati.

- *M*: è il linguaggio alla base di Power Query Editor. Esso è usato per effettuare trasformazioni molto complesse che vanno al di là delle potenzialità dell'interfaccia grafica dell'editor; tali operazioni avvengono utilizzando il linguaggio M nell'editor avanzato.
- *T-SQL*: permette di scrivere query native per estrarre, filtrare e combinare i dati direttamente dal server di origine prima ancora che vengano caricati nel report. Questa pratica ottimizza le prestazioni riducendo il carico di lavoro di Power BI e velocizzando i tempi di aggiornamento.
- *Python e R*: vengono usati per creare script sia per la preparazione dei dati sia per la creazione di visualizzazioni avanzate. Questi script possono essere realizzati tramite un IDE esterno o direttamente all'interno di Power BI. I Data Scientist utilizzano questi linguaggi poiché mettono a disposizione librerie dedicate per eseguire analisi statistiche complesse e modelli di machine learning.

CAPITOLO 3

Descrizione dei dataset di partenza

In questo capitolo vengono descritti i due dataset impiegati per l'analisi condotta tramite Power BI, evidenziandone le caratteristiche principali e la struttura. Successivamente, verrà introdotto il processo di ETL, utilizzato per la selezione e preparazione dei dati nelle sue tre fasi: estrazione, trasformazione e caricamento. Infine, verrà spiegato come questo processo sia stato applicato ai due dataset di interesse, passo necessario per la loro integrazione in Power BI.

3.1 Dataset utilizzati

L'analisi condotta mediante Power BI è stata realizzata tramite due dataset pubblici presenti sulla piattaforma web Kaggle (`www.kaggle.com`). Questi dataset sono stati scelti perché forniscono diverse informazioni riguardanti il mondo del lavoro relativo all'Intelligenza Artificiale e degli indici che ne rappresentano lo sviluppo in vari Paesi.

Il primo, *Global AI Job Market and Salary Trends 2025*, contiene informazioni dettagliate su posizioni lavorative, stipendi e tendenze di mercato nel campo dell'IA e del machine learning in diversi Paesi, livelli di esperienza e dimensioni aziendali. Esso è progettato per supportare diversi casi d'uso in svariati ambiti. Ad esempio, può essere utilizzato per creare modelli di previsione salariale, capire quali fattori influenzano la retribuzione e confrontare i guadagni tra le diverse città o regioni. Inoltre, permette di analizzare le tendenze del mercato del lavoro, monitorando la sua evoluzione, individuando i nuovi ruoli e le competenze più richieste, e studiando la diffusione del lavoro a distanza.

Queste informazioni sono utili anche per chi vuole pianificare la propria carriera, poiché aiutano a comprendere i requisiti necessari per ogni posizione e a confrontare le opportunità nei vari Paesi. Dal punto di vista aziendale e della Business Intelligence, il dataset consente di analizzare i modelli di assunzione, individuare la mancanza di competenze e studiare la concorrenza. Infine, si presta a studi geografici per valutare il rapporto tra lo stipendio e il costo della vita, oltre a misurare la maturità dei mercati locali.

Il secondo, *AI Global Index*, include l'indice globale sull'AI e sette indicatori che lo influenzano in 62 Paesi, oltre a informazioni generali su di essi. Esso permette di confrontare le nazioni in base al loro livello di investimento, innovazione e implementazione dell'Intelligenza Artificiale.

Tutti questi sette indicatori sono stati calcolati da Tortoise Media ponderando e riassumendo altri 143 sotto-indicatori. Il dataset può essere utilizzato per esercitarsi nella pulizia e

nella visualizzazione dei dati, nell'individuazione di correlazioni tra gli indici e nel machine learning (classificazione, regressione, clustering).

La scelta di utilizzare questi due dataset è dovuta al fatto che il primo dà una visione microscopica sul mondo del lavoro nel settore AI, mentre il secondo offre una visione macroscopica grazie all'utilizzo di diversi indici; tale combinazione permette di ottenere una base di dati completa e robusta per l'analisi.

3.1.1 Dataset 1: Global AI Job Market and Salary Trends 2025

Il dataset *Global AI Job Market and Salary Trends 2025* fornisce un'analisi approfondita del mercato del lavoro nel settore dell'Intelligenza Artificiale, con oltre 15.000 offerte di lavoro reali raccolte dalle principali piattaforme di ricerca lavoro a livello mondiale. Esso include informazioni dettagliate su stipendi, requisiti professionali, approfondimenti aziendali e tendenze geografiche. Le colonne da cui è composto sono:

1. *job_id*: identificativo univoco per ogni annuncio di lavoro;
2. *job_title*: titolo professionale nel settore AI;
3. *salary_usd*: stipendio annuo in dollari;
4. *salary_currency*: valuta originale dello stipendio;
5. *salary_local*: stipendio annuo nella valuta originale;
6. *experience_level*: livello di esperienza, come EN (Entry), MI (Mid), SE (Senior) ed EX (Executive);
7. *employment_type*: tipologia di impiego, come FT (Full-time), PT (Part-time), CT (Contract) e FL (Freelance);
8. *company_location*: Paese in cui ha sede l'azienda;
9. *company_size*: dimensione dell'azienda, come S (Small), M (Medium) e L (Large);
10. *employee_residence*: Paese di residenza del dipendente;
11. *remote_ratio*: rapporto che indica se un impiegato lavora o meno da remoto; può assumere tre valori, come 0 (No Remote), 50 (Hybrid) e 100 (Fully remote);
12. *required_skills*: elenco delle cinque abilità più richieste;
13. *education_required*: requisito minimo di istruzione, come Associate, Bachelor, Master e PhD;
14. *years_experience*: numero di anni di esperienza richiesti;
15. *industry*: settore industriale dell'azienda;
16. *posting_date*: data di pubblicazione dell'annuncio di lavoro (dal 2024-01-01 fino al 2025-04-30);
17. *application_deadline*: termine ultimo per la presentazione delle domande di lavoro (dal 2024-01-16 fino al 2025-07-11);
18. *job_description_length*: numero di caratteri della descrizione del lavoro;
19. *benefits_score*: punteggio numerico del pacchetto dei benefici dati dall'azienda (1-10);
20. *company_name*: nome dell'azienda.

3.1.2 Dataset 2: AI Global Index

Il dataset *AI Global Index* registra 62 Paesi, fornendo per ciascuno di essi un indice globale sull'AI e le informazioni ad esso relative. Inoltre, esso contiene altri sette indici che possono essere raggruppati in categorie relative all'implementazione (Talent, Infrastructure and Operating Environment), all'innovazione (Research e Development) e agli investimenti (Government Strategy e Commercial); tutti gli indici assumono valori tra 0 e 100. Le colonne che formano questo dataset sono:

1. *Country*: Paese a cui saranno associati i vari indici;
2. *Talent*: indicatore che si concentra sulla disponibilità di professionisti qualificati per la fornitura di soluzioni di Intelligenza Artificiale;
3. *Infrastructure*: indicatore che si concentra sull'affidabilità e sulla portata delle infrastrutture di accesso, dall'elettricità e internet alle capacità di supercalcolo;
4. *Operating Environment*: indicatore che si concentra sul contesto normativo e sull'opinione pubblica in merito all'Intelligenza Artificiale;
5. *Research*: indicatore che si concentra sull'ampiezza della ricerca specialistica e sul numero di ricercatori, analizzando la quantità di pubblicazioni e citazioni in riviste accademiche autorevoli;
6. *Development*: indicatore che si concentra sullo sviluppo di piattaforme e algoritmi fondamentali su cui si basano i progetti innovativi di Intelligenza Artificiale;
7. *Government Strategy*: indicatore che si concentra sulla profondità dell'impegno del governo nazionale nei confronti dell'Intelligenza Artificiale, analizzando gli impegni di spesa e le strategie nazionali;
8. *Commercial*: indicatore che si concentra sul livello di attività delle startup, sugli investimenti e sulle iniziative imprenditoriali basate sull'Intelligenza Artificiale;
9. *Total score*: indicatore globale sull'AI che confronta le nazioni in base al loro livello di investimento, innovazione e implementazione;
10. *Region*: regione in cui si trova il Paese, ad esempio Europe, Asia-Pacific, ecc.;
11. *Cluster*: raggruppamento dei Paesi in categorie in base al loro sviluppo nel settore AI, come Nascent, Waking up, Rising stars, Traditional champions e Power players;
12. *Income group*: tipo di fascia di reddito del Paese, ad esempio High, Upper middle e Lower middle;
13. *Political regime*: regime politico del Paese.

3.2 Processo di ETL

ETL (estrazione, trasformazione, caricamento) è un processo di integrazione dei dati che combina, pulisce e organizza i dati provenienti da più fonti in un unico set di dati uniforme. Successivamente, esso carica quindi i dati in un data warehouse, data lake o altro sistema di destinazione.

Le pipeline di dati ETL forniscono le basi per l'analisi dei dati e i flussi di lavoro di machine learning. Attraverso una serie regole di business, l'ETL ripulisce e organizza i dati

in modo da soddisfare esigenze specifiche di Business Intelligence. Essa può anche affrontare attività di analytics più avanzate, che possono migliorare i processi di back-end e le esperienze degli utenti finali. Le pipeline ETL vengono spesso utilizzate dalle organizzazioni per:

- estrarre i dati dai sistemi legacy;
- pulire i dati per migliorarne la qualità e stabilirne la coerenza;
- caricare i dati in un database di destinazione.

Le soluzioni ETL migliorano la qualità eseguendo la pulizia dei dati prima di caricare i dati su un repository diverso. Essendo un'operazione batch che richiede molto tempo, l'ETL è consigliato più spesso per la creazione di repository di dati di destinazione più piccoli, che richiedono aggiornamenti meno frequenti. Altri metodi di integrazione dei dati, tra cui ELT (Extract, Load, Transform), CDC (Change Data Capture) e la virtualizzazione dei dati, sono utilizzati per integrare volumi di dati sempre più grandi che cambiano i flussi di dati in tempo reale.

In passato, le organizzazioni scrivevano il proprio codice ETL. Oggi sono disponibili molti strumenti ETL open source e commerciali, così come servizi basati sul cloud, tra cui scegliere. Le capacità tipiche di questi prodotti includono:

- *Automazione completa e facilità d'uso*: i principali strumenti ETL automatizzano l'intero flusso di dati, dalle fonti dei dati al data warehouse di destinazione. Questa automazione evita ai data engineer le noiose attività di spostamento e formattazione dei dati, per risultati più rapidi e per effettuare operazioni in modo più efficiente.
- *Un'interfaccia visiva drag-and-drop*: questa funzionalità può essere utilizzata per specificare regole e flussi di dati.
- *Supporto per la gestione di dati complessi*: questa funzionalità include l'assistenza per calcoli complessi, integrazioni di dati e manipolazioni di stringhe.
- *Sicurezza e conformità*: i migliori strumenti ETL crittografano i dati sia in movimento che a riposo e hanno conseguito certificazioni di conformità alle normative di settore o governative, tra cui HIPAA e GDPR.

3.2.1 Introduzione

Il processo di ETL si articola in tre fasi (Figura 3.1), ognuna relativa ad uno specifico compito di elaborazione:

- *Extract (Estrazione)*: i dati non elaborati vengono copiati o esportati dalle posizioni di origine a un'area di staging. I team di gestione dei dati possono estrarli da fonti diverse, strutturate o non strutturate. Questi tipi di dati includono: server SQL o NoSQL, sistemi CRM ed ERP, JSON e XML, flat-file database e pagine web.
- *Transform (Trasformazione)*: nell'area di staging i dati non elaborati vengono trasformati e consolidati per il caso d'uso analitico previsto per essi. Durante questa fase, le informazioni grezze vengono inizialmente filtrate, pulite, aggregate e deduplicate, per poi essere sottoposte a rigorosi processi di convalida e autenticazione. Successivamente, si procede all'esecuzione di calcoli, traduzioni o riepiloghi basati sui dati d'origine; questa attività comprende operazioni pratiche, come la modifica delle intestazioni di righe e colonne per garantire la coerenza strutturale, la conversione di valute o altre unità di misura e la manipolazione di stringhe di testo. Al contempo, vengono eseguiti controlli

di audit per calcolare le metriche necessarie e garantire la qualità complessiva del dato. In quest'ottica, i dati regolati dalle autorità di settore o governative vengono opportunamente rimossi, crittografati o protetti. Infine, l'intero set di dati viene formattato in tabelle, eventualmente tra loro collegate, in modo da corrispondere perfettamente allo schema strutturale del data warehouse di destinazione.

- *Load (Caricamento)*: in quest'ultimo passaggio, i dati trasformati vengono spostati dall'area di staging a un data warehouse di destinazione. In genere, questo processo comporta un caricamento iniziale di tutti i dati, seguito dal caricamento periodico delle modifiche incrementali dei dati e, meno spesso, da aggiornamenti completi per cancellare e sostituire i dati nel data warehouse.

Per la maggior parte delle organizzazioni che utilizzano ETL, il processo è automatizzato, ben definito, continuo e basato su batch. In genere, la fase di caricamento dei dati avviene durante le ore di riposo, quando il traffico sui sistemi di origine e sul data warehouse è al minimo.

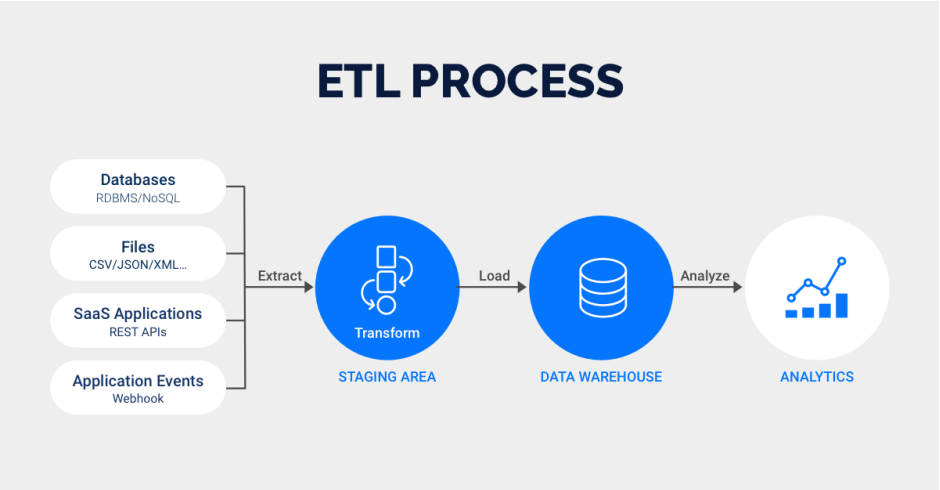


Figura 3.1: Fasi del processo di ETL

3.2.2 Attività di ETL effettuate sul primo dataset

Nel dataset *Global AI Job Market and Salary Trends 2025*, descritto nella Sezione 3.1.1, è stata eseguita un'attività di ETL per prepararlo all'analisi.

Extract

La fase di estrazione è stata eseguita scaricando il dataset dalla piattaforma Kaggle in formato `.csv`. In Power BI Desktop, per eseguire questa operazione, è necessario:

- avviare l'applicazione e selezionare *Home* > *Recupera Dati* > *Testo/CSV* (Figura 3.2);
- selezionare il file `.csv` sul dispositivo locale;
- premere il pulsante *Apri*, in modo che venga visualizzata un'anteprima del contenuto del file (Figura 3.3). Esaminando l'anteprima Power BI tenta di identificare:
  - il carattere separatore dei campi (virgola, punto e virgola o tabulazione);
  - l'eventuale presenza di intestazioni nella riga iniziale del documento;

- la tipologia di dato associata a ciascuna colonna (valori testuali, numerici, date, ecc.).

▶ scegliere tra due modalità distinte:

- selezionare l'opzione *Carica* per inserire direttamente i dati all'interno del modello;
- selezionare l'opzione *Trasforma dati* per reindirizzare il flusso di lavoro verso il Power Query Editor; questa seconda opzione permette di effettuare operazioni preliminari di pulizia, modellazione e raffinamento del dataset prima del suo effettivo caricamento.

Durante la fase di estrazione, Power BI valuta se la prima riga del file contiene i nomi delle colonne e, in caso positivo, la promuove automaticamente come intestazione della tabella che verrà generata.

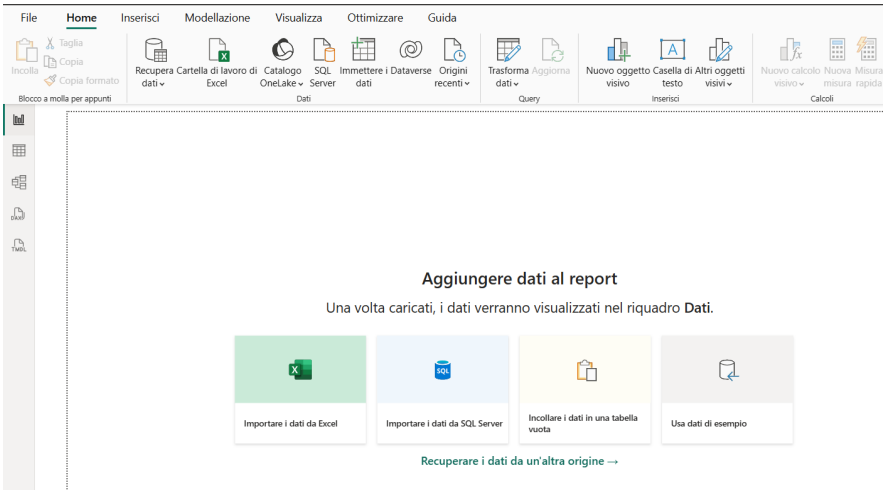
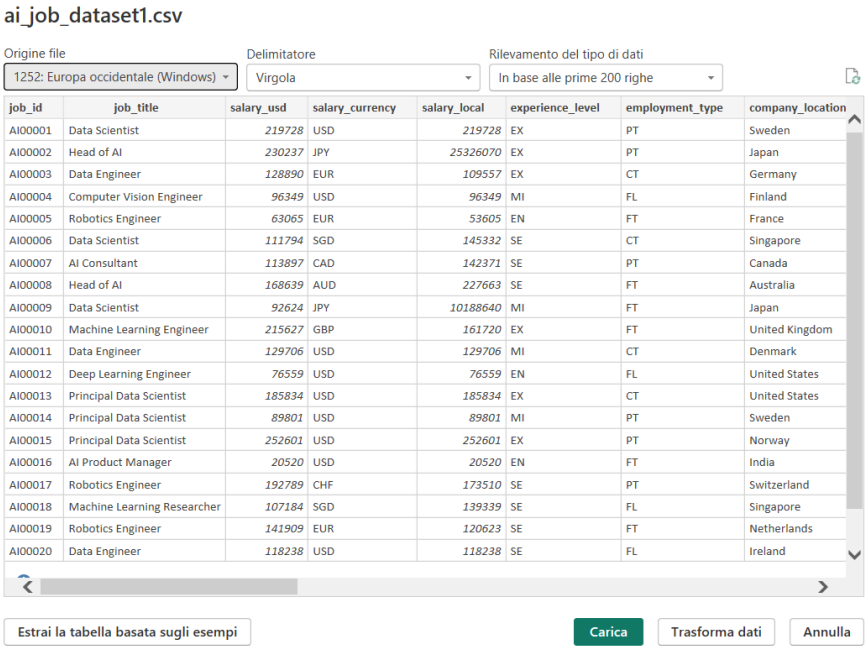


Figura 3.2: Importazione dei dati in Power BI



job_id	job_title	salary_usd	salary_currency	salary_local	experience_level	employment_type	company_location
AI00001	Data Scientist	219728	USD	219728	EX	PT	Sweden
AI00002	Head of AI	230237	JPY	25326070	EX	PT	Japan
AI00003	Data Engineer	128890	EUR	109557	EX	CT	Germany
AI00004	Computer Vision Engineer	96349	USD	96349	MI	FL	Finland
AI00005	Robotics Engineer	63065	EUR	53605	EN	FT	France
AI00006	Data Scientist	111794	SGD	145332	SE	CT	Singapore
AI00007	AI Consultant	113897	CAD	142371	SE	PT	Canada
AI00008	Head of AI	168639	AUD	227663	SE	FT	Australia
AI00009	Data Scientist	92624	JPY	10188640	MI	FT	Japan
AI00010	Machine Learning Engineer	215627	GBP	161720	EX	FT	United Kingdom
AI00011	Data Engineer	129706	USD	129706	MI	CT	Denmark
AI00012	Deep Learning Engineer	76559	USD	76559	EN	FL	United States
AI00013	Principal Data Scientist	185834	USD	185834	EX	CT	United States
AI00014	Principal Data Scientist	89801	USD	89801	MI	PT	Sweden
AI00015	Principal Data Scientist	252601	USD	252601	EX	PT	Norway
AI00016	AI Product Manager	20520	USD	20520	EN	FT	India
AI00017	Robotics Engineer	192789	CHF	173510	SE	PT	Switzerland
AI00018	Machine Learning Researcher	107184	SGD	139339	SE	FL	Singapore
AI00019	Robotics Engineer	141909	EUR	120623	SE	FT	Netherlands
AI00020	Data Engineer	118238	USD	118238	SE	FL	Ireland

Figura 3.3: Anteprima del dataset da importare e le opzioni di caricamento

Transform

La fase di trasformazione dei dati in Power BI avviene selezionando l'opzione *Trasforma dati* in modo da poter utilizzare il Power Query Editor per renderli coerenti e pronti per l'analisi. Nel caso del dataset *Global AI Job Market and Salary Trends 2025* non sono state necessarie operazioni di pulizia preliminare, data l'assenza di valori nulli, errati o di righe duplicate.

Durante la fase di trasformazione del primo dataset è stato necessario modificare il tipo della colonna *benefits_score*, poiché quando sono stati importati i dati, Power BI aveva interpretato tale colonna come intera, sebbene essa contenesse originariamente valori decimali. Questa operazione di conversione del tipo da `Int64.Type` in `number` viene visualizzato nel Listato 3.1.

```
#"Modificato tipo" =
Table.TransformColumnTypes(#"Intestazioni alzate di livello",
 {{"job_id", type text}, {"job_title", type text},
 {"salary_usd", Int64.Type},{"salary_currency", type text},
 {"salary_local", Int64.Type}, {"experience_level", type text},
 {"employment_type", type text},
 {"company_location", type text},
 {"company_size", type text},
 {"employee_residence", type text},
 {"remote_ratio", Int64.Type}, {"required_skills", type text},
 {"education_required", type text},
 {"years_experience", Int64.Type},
 {"industry", type text}, {"posting_date", type date},
 {"application_deadline", type date},
 {"job_description_length", Int64.Type},
 {"benefits_score", type number},
 {"company_name", type text}},"en-US"),
```

Listing 3.1: Modifica del tipo di dato nella colonna *benefits_score*

Sebbene l'analisi preliminare del dataset non abbia evidenziato la presenza di valori nulli, si è ritenuto opportuno implementare una logica di gestione dei dati mancanti. Questa misura precauzionale mira a garantire la robustezza e la replicabilità del processo di ETL, prevenendo potenziali anomalie nel modello dei dati qualora il dataset originario venisse aggiornato in futuro con record incompleti. Gli eventuali dati numerici mancanti sono stati sostituiti inserendo al loro posto lo zero, come mostrato nel Listato 3.2.

```
#"Colonne numeriche" = {"salary_usd","salary_local",
"remote_ratio","years_experience","job_description_length",
"benefits_score"},
#"Rimpiazza valori nulli" =
Table.ReplaceValue( #"Modificato tipo", null, 0,Replacer.
 ReplaceValue, #"Colonne numeriche"),
```

Listing 3.2: Sostituzione di dati numerici mancanti nel primo dataset

Gli eventuali dati di testo mancanti, invece, sono stati sostituiti inserendo al loro posto la stringa "Unknown", come mostrato nel Listato 3.3.

```
#"Colonne di testo" =
{"job_id","job_title","salary_currency",
"experience_level","employment_type","company_location",
"company_size","employee_residence","required_skills",
"education_required","industry","company_name"},
#"Rimpiazza testo nullo" =
```

```
7     Table.ReplaceValue(#"Rimpiazza valori nulli", null,
8     "Unknown", Replacer.ReplaceValue, #"Colonne di testo"),
```

Listing 3.3: Sostituzione di dati testuali mancanti nel primo dataset

Un'ulteriore operazione di trasformazione ha riguardato la traduzione delle abbreviazioni presenti nelle colonne *experience_level* ed *employment_type*. I codici presenti inizialmente sono stati sostituiti con i rispettivi valori estesi, al fine di migliorare la leggibilità dei dati e ottimizzare l'esperienza degli utenti durante la fase di visualizzazione.

Nello specifico, i codici della colonna *experience_level* sono stati mappati come riassunto nella Tabella 3.1.

Valore originale	Valore esteso
EN	Entry level
MI	Mid level
SE	Senior level
EX	Executive level

Tabella 3.1: Mappatura dei codici della colonna *experience_level*

Questa mappatura è stata realizzata usando il linguaggio M, per effettuare la sostituzione in un unico passaggio, come illustrato nel Listato 3.4.

```
1     #"mapping experience_level" = [
2         #"EN" = "Entry level",
3         #"MI" = "Mid level",
4         #"SE" = "Senior level",
5         #"EX" = "Executive level"
6     ],
7
8     #"sostituto experience_level" = Table.TransformColumns(
9         #"Rimpiazza testo nullo",
10        {
11            {
12                "experience_level",
13                each Record.FieldOrDefault(#"mapping
      experience_level", _, _),
14                type text
15            }
16        }
17    ),
```

Listing 3.4: Sostituzione dei codici della colonna *experience_level*

Invece, la traduzione dei codici usati nella colonna *employment_type* è stata riassunta nella Tabella 3.2.

Valore originale	Valore esteso
FT	Full-time
PT	Part-time
CT	Contract
FL	Freelance

Tabella 3.2: Mappatura dei codici della colonna *employment_type*

Analogamente a quanto effettuato per *experience_level*, anche per la colonna *employment_type* è stata predisposta una struttura di mappatura per eseguire la sostituzione dei valori in un unico passaggio, come mostrato nel Listato 3.5.

```
1     #"mapping employment_type" = [
2          #"FT" = "Full-time",
3          #"PT" = "Part-time",
4          #"CT" = "Contract",
5          #"FL" = "Freelance"
6      ],
7
8      #"sostituto employment_type" = Table.TransformColumns(
9          #"sostituto experience_level",
10         {
11             {
12                 "employment_type",
13                 each Record.FieldOrDefault(#"mapping
       employment_type", _, _),
14                 type text
15             }
16         }
17     ),
```

Listing 3.5: Sostituzione dei codici della colonna *employment_type*

Anche se da un primo controllo i dati nelle colonne *job_title* ed *education_required* sembravano già scritti in modo coerente, è stato deciso, comunque, di applicare la funzione `Text.Lower` per convertire tutto il testo in minuscolo.

Si tratta di una scelta utile a evitare problemi nel caso in cui i futuri aggiornamenti del dataset presentino inserimenti manuali non standardizzati. Queste due colonne, infatti, contengono parole estese e sono le più importanti per i filtraggi all'interno delle dashboard. Se il testo non fosse perfettamente identico, si rischierebbe lo sdoppiamento delle categorie nei grafici. La trasformazione è riportata nel Listato 3.6.

```
1   #"job_title in minuscolo" = Table.TransformColumns(#"sostituto
      employment_type",{{"job_title", Text.Lower, type text}}),
2
3   #"education_required in
      minuscolo" = Table.TransformColumns(#"job_title in minuscolo
      ",{{"education_required", Text.Lower, type text}}),
```

Listing 3.6: Normalizzazione dei testi nelle colonne *job_title* ed *education_required*

La colonna *required_skills*, essendo formata da liste di valori testuali, è stata eliminata dalla tabella principale ed è stata utilizzata per creare una nuova tabella chiamata *Skill_Analysis* (trattata in seguito) per evitare che l'eventuale separazione dei valori delle liste generasse dei record ridondanti nella tabella principale. L'operazione di rimozione della colonna è descritta nel Listato 3.7.

```
1     #"Rimosse colonne" = Table.RemoveColumns(#"education_required
       in minuscolo",{"required_skills"}),
```

Listing 3.7: Rimozione della colonna *required_skills*

Al fine di estrarre informazioni sul tempo di candidatura concesso dalle aziende, è stata creata la colonna *days_to_apply*, calcolata come la differenza matematica tra la data di scadenza dell'annuncio (*application_deadline*) e la data di pubblicazione dello stesso (*posting_date*) convertendo, poi, il risultato in un valore intero. L'operazione è illustrata nel Listato 3.8.

```
1      #"aggiunta days_to_apply"= Table.AddColumn(#"Rimosse colonne
       ","days_to_apply", each Duration.Days([application_deadline]
       - [posting_date]), Int64.Type),
```

Listing 3.8: Aggiunta della colonna *days_to_apply*

Un'altra operazione effettuata su questo dataset riguarda la standardizzazione delle intestazioni, in modo da renderle più coerenti con quelle del secondo dataset e facilitarne i confronti. I nomi delle colonne, dopo il processo di ridenominazione, sono:

- Job ID;
- Job Title;
- Salary USD;
- Salary Currency;
- Salary Local;
- Experience Level;
- Employment Type;
- Company Location;
- Company Size;
- Employee Residence;
- Remote Ratio;
- Education Required;
- Years Experience;
- Industry;
- Posting Date;
- Application Deadline;
- Job Description Length;
- Benefit Score;
- Company Name;
- Days To Apply.

Questa operazione viene descritta nel Listato 3.9.

```
1      #"Ridenominazione colonne" = Table.RenameColumns(#"aggiunta
       days_to_apply",{{"job_id","Job ID"},
2      {"job_title","Job Title"},{"salary_usd","Salary USD"},{"
       salary_currency","Salary Currency"},
3      {"salary_local","Salary Local"},{"experience_level","
       Experience Level"},
4      {"employment_type","Employment Type"},{"company_location","
       Company Location"},
```

```
5     {"company_size","Company Size"},{"employee_residence","
      Employee Residence"},
6     {"remote_ratio","Remote Ratio"}, {"education_required","
      Education Required"},
7     {"years_experience","Years Experience"},{"industry","Industry
      "},
8     {"posting_date","Posting Date"},{"application_deadline","
      Application Deadline"},
9     {"job_description_length","Job Description Length"},{"
      benefits_score","Benefit Score"},
10    {"company_name","Company Name"},{"days_to_apply","Days To
      Apply"} }),
```

Listing 3.9: Ridenominazione delle intestazioni delle colonne

Nei grafici l'ordinamento dei dati testuali viene fatto da Power BI in base all'ordine alfabetico, per questo è stato necessario creare una colonna condizionale chiamata *Order of Experience Level* in modo da ordinare i valori di *Experience Level* in base al grado di esperienza (Listato 3.10).

```
1       #"Order of Experience Level" = Table.AddColumn(#"
      Ridenominazione colonne", "Order of Experience Level",
2       each if [Experience Level] = "Entry level" then 1
3       else if [Experience Level] = "Mid level" then 2
4       else if [Experience Level] = "Senior level" then 3
5       else if [Experience Level] = "Executive level" then 4
6       else null),
```

Listing 3.10: Aggiunta della colonna *Order of Experience Level*

Per stabilire un'eventuale relazione con il secondo dataset i valori della colonna *Company Location* devono essere puliti eliminando eventuali caratteri invisibili o spazi aggiuntivi, come viene mostrato nel Listato 3.11.

```
1     #"Pulizia nomi Paesi"= Table.TransformColumns(#"Order of
      Experience Level",{{"Company Location",each Text.Trim(Text.
      Clean(_)),type text}}),
```

Listing 3.11: Pulizia della colonna *Company Location*

Infine, per migliorare la visibilità dei dati della colonna *Remote Ratio* è stata creata una nuova colonna condizionale chiamata *Work Mode* che associa ad ogni numero il corrispondente significato, come viene illustrato nel Listato 3.12.

```
1     #"Colonna condizionale Work
      Mode" = Table.AddColumn(#"Pulizia nomi Paesi", "Work Mode",
2     each if [Remote Ratio] = 0 then "No Remote"
3     else if [Remote Ratio] = 50 then "Hybrid"
4     else if [Remote Ratio] = 100 then "Fully Remote"
5     else null)
```

Listing 3.12: Aggiunta della colonna *Work Mode*

Come accennato in precedenza, la tabella *Skill_Analysis* è stata creata come un flusso separato a partire dal dataset iniziale, al fine di isolare e analizzare le competenze richieste senza appesantire la tabella principale. Il processo ha previsto l'estrazione della colonna *job_id*, necessaria per mantenere il legame relazionale, e della colonna *required_skills*. Successivamente, i valori testuali racchiusi all'interno delle liste sono stati separati in modo che ogni singola competenza occupasse una riga autonoma. Questa operazione ha permesso di

strutturare una relazione di tipo "uno a molti", fondamentale per evitare la duplicazione di record del dataset. Le *required_skills* sono state separate utilizzando la virgola per generare nuove righe; successivamente sono stati eliminati gli spazi in eccesso ed è stata effettuata la ridenominazione delle intestazioni delle colonne, rispettivamente in: Job ID e Required Skills. Le trasformazioni applicate sono descritte nel Listato 3.13.

```
#"Rimosse colonne" = Table.RemoveColumns(#"Modificato tipo",{
"job_title", "salary_usd", "salary_currency", "salary_local"
, "experience_level", "employment_type", "company_location",
 "company_size", "employee_residence", "remote_ratio", "
education_required", "years_experience", "industry", "
posting_date", "application_deadline", "
job_description_length", "benefits_score", "company_name"}),

#"divisione required_skill" = Table.ExpandListColumn(Table.
TransformColumns(#"Rimosse colonne", {{"required_skills",
Splitter.SplitTextByDelimiter(","), type list}}), "
required_skills"),

#"rimosso spazi required_skill" = Table.TransformColumns(#"
divisione required_skill", {{"required_skills", Text.Trim,
type text}}),

#"required_skill in minuscolo" = Table.TransformColumns(#"
rimosso spazi required_skill", {{"required_skills", Text.
Lower, type text}}),

#"Ridenominazione colonne" = Table.RenameColumns(#"
required_skill in minuscolo",{{"job_id","Job ID"},{"
required_skills","Required Skills"} })
```

Listing 3.13: Operazioni sulla tabella *Skill_Analysis*

Load

Una volta completata la fase di trasformazione dei dati in *Power Query Editor*, il caricamento viene effettuato tramite l'opzione *Chiudi e applica* per importare il dataset all'interno di Power BI. Il primo dataset è stato caricato su Power BI dove sono state eseguite le seguenti operazioni:

- *relazioni tra tabelle*: poiché il primo dataset utilizza come chiave primaria il campo *Job ID*, la stessa colonna utilizzata nella creazione della tabella *Skill_Analysis*, è stata creata una relazione "uno a molti" tra i due utilizzando proprio quest'identificatore come chiave esterna;
- *creazione di misure*: all'interno del modello Power BI sono state definite diverse misure e colonne calcolate tramite DAX, utili per l'analisi e la realizzazione di report e dashboard.

3.2.3 Attività di ETL effettuate sul secondo dataset

Nel dataset *AI Global Index*, descritto nella Sezione 3.1.2, è stata eseguita un'attività di ETL per prepararlo all'analisi, come fatto in precedenza anche per il primo dataset.

Extract

La fase di estrazione è stata eseguita scaricando il dataset dalla piattaforma Kaggle in formato `.csv`.

Per eseguire questa fase è stato seguito lo stesso procedimento descritto nella Sezione 3.2.2.

Transform

Nel caso del dataset *AI Global Index* non sono state necessarie operazioni di pulizia preliminare, data l'assenza di valori nulli, errati o di righe duplicate. Di default, durante l'importazione in Power BI, i valori numerici vengono interpretati come numeri interi. Essendo questo dataset costituito da molti indici, quindi valori decimali, è stato necessario convertire il tipo di tali indici, come mostrato nel Listato 3.14.

```
1  #"Modificato tipo" = Table.TransformColumnTypes(#"
   Intestazioni alzate di livello",
2  {{"Country", type text}, {"Talent", type number}, {"
   Infrastructure", type number},
3   {"Operating Environment", type number}, {"Research", type
   number},
4   {"Development", type number}, {"Government Strategy",type
   number}, {"Commercial", type number},
5   {"Total score",type number}, {"Region", type text}, {"
   Cluster", type text},
6   {"Income group", type text}, {"Political regime", type text
   }},"en-US"),
```

Listing 3.14: Modifica dei tipi di dati nel secondo dataset

Sebbene l'analisi preliminare del dataset non abbia evidenziato la presenza di valori nulli, si è ritenuto opportuno implementare una logica di gestione dei dati mancanti. Questa misura precauzionale mira a garantire la robustezza e la replicabilità del processo di ETL, prevenendo potenziali anomalie nel modello dati qualora il dataset originario venisse aggiornato in futuro con record incompleti. Gli eventuali dati numerici mancanti sono stati sostituiti inserendo al loro posto lo zero, come mostrato nel Listato 3.15.

```
1  #"Colonne numeriche"= {"Talent","Infrastructure","Operating
   Environment","Research","Development","Government Strategy",
   "Commercial","Total score"},
2
3  #"Rimpiazza valori nulli"=Table.ReplaceValue( #"Modificato
   tipo", null, 0,Replacer.ReplaceValue, #"Colonne numeriche"),
```

Listing 3.15: Sostituzione di dati numerici mancanti nel secondo dataset

Gli eventuali dati di testo mancanti, invece, sono stati sostituiti inserendo al loro posto la stringa "Unknown", come mostrato nel Listato 3.16.

```
1  #"Colonne di testo"={"Country","Region","Cluster","Income
   group","Political regime"},
2
3  #"Rimpiazza testo
   nullo"=Table.ReplaceValue( #"Rimpiazza valori nulli", null,
   "Unknown",Replacer.ReplaceValue, #"Colonne di testo"),
```

Listing 3.16: Sostituzione di dati testuali mancanti nel secondo dataset

Il primo e il secondo dataset hanno in comune una colonna in cui vengono inseriti i Paesi; questa verrà poi utilizzata per creare una relazione tra i due. A tal fine, è necessario uniformare i nomi dei Paesi nel secondo dataset, come illustrato nel Listato 3.17.

```
#"Pulizia nomi
  Paesi" = Table.TransformColumns(#"Rimpiazza testo nullo",{{"
  Country",each Text.Trim(Text.Clean(_)),type text}}),

#"Mapping nomi paesi"=[
        #"The Netherlands"="Netherlands",
        #"United States of America"="United States"
    ],

#"Cambia nomi paesi"=Table.TransformColumns(
        #"Pulizia nomi Paesi",
        {
            {
                "Country",
                each Record.FieldOrDefault(#"Mapping nomi paesi"
  , _, _),
                type text
            }
        }
    ),
```

Listing 3.17: Modifica dei nomi dei Paesi nel secondo dataset

Infine, per la realizzazione dei grafici è stato necessario creare una copia del secondo dataset in cui viene eseguita un'operazione di Unpivot solo per gli indici. Con questa operazione vengono create due nuove colonne *Index* e *Value* in cui nella prima viene specificato il nome dell'indice mentre nella seconda il valore effettivo. In questo modo, vengono messe a disposizione due versioni del secondo dataset in modo da poter attingere ai campi dell'una o dell'altra tabella in base alle specifiche necessità di visualizzazione. L'operazione di Unpivot viene riportata nel Listato 3.18.

```
#"Trasformate tramite UnPivot solo le colonne selezionate" =
Table.Unpivot(#"Cambia nomi paesi", {"Talent", "Infrastructure"
  , "Operating Environment", "Research", "Development", "
  Government Strategy", "Commercial"}, "Index", "Value"),
```

Listing 3.18: Operazione di Unpivot sulle colonne selezionate

Load

Una volta completate le operazioni di trasformazione sul secondo dataset, esso è stato caricato in Power BI. Durante questa fase, come anche per il primo dataset, sono state create delle relazioni e misure DAX.

Le misure DAX, insieme a quelle del primo dataset, sono state raggruppate all'interno di una tabella dedicata chiamata *_Misure*, in modo da separare nettamente la logica di calcolo dalla struttura dei dati grezzi, migliorando la manutenibilità e la leggibilità dell'intero modello.

Per quanto riguarda le relazioni, ne è stata creata una con molteplicità "uno a molti" tra *AI Global Index* e *Global AI Job Market and Salary Trends 2025* sfruttando la chiave primaria *Country* del secondo dataset e come chiave esterna la colonna *Company Location* del primo. Tale collegamento permette di confrontare i dati su stipendi e offerte di lavoro con i vari

indici dell'IA di ogni specifico Paese. Inoltre, per completare il modello, è stata aggiunta una seconda relazione "uno a molti" tra *AI Global Index* e la sua versione in cui è stata eseguita l'operazione di Unpivot (*AI_index_UNIPIVOT*). Questa relazione si rende necessaria affinché i visual creati si aggiornino correttamente in base alle selezioni e ai filtri applicati nel resto del report.

Tutte le relazioni create nelle fasi di ETL sia del primo che del secondo dataset sono rappresentate nella Figura 3.4.

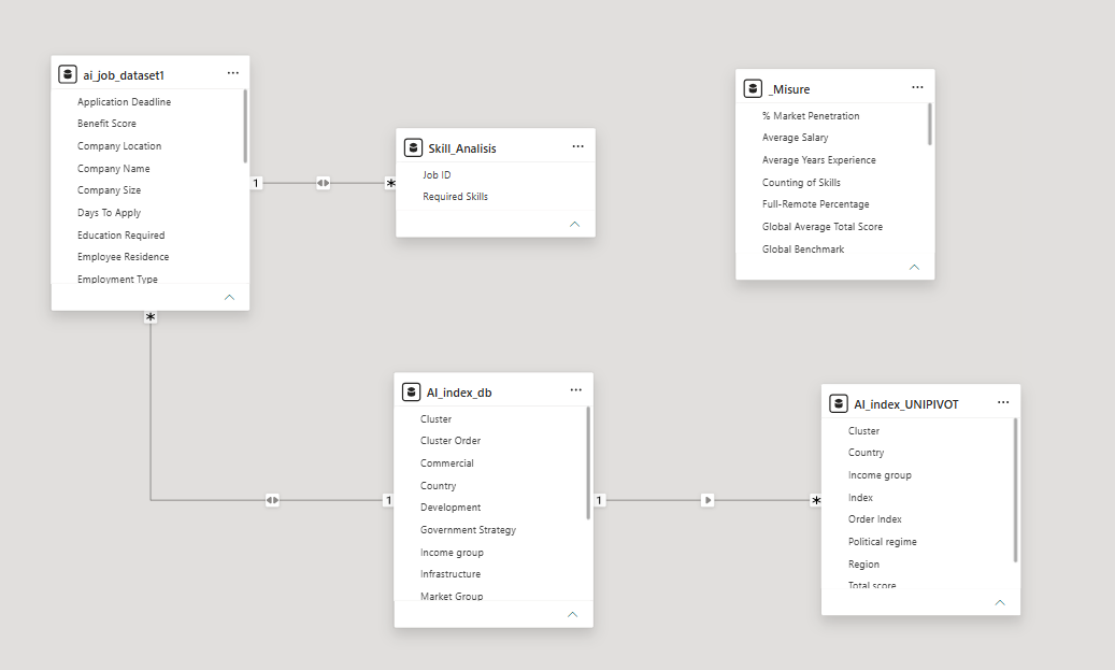


Figura 3.4: Vista delle relazioni del modello analizzato

CAPITOLO 4

Progettazione e realizzazione delle dashboard di base

In questo capitolo vengono presentati i visual e i grafici offerti da Power BI, per avere una panoramica sugli strumenti utilizzati per la creazione dei report. Successivamente, vengono elencate le misure DAX implementate, fondamentali per estrarre indicatori e metriche che aggiungono valore alle informazioni di partenza. Questo capitolo si conclude con l'analisi delle prime dashboard di base, le quali sono state sviluppate a partire dai dati raccolti nel dataset "Global AI Job Market and Salary Trends 2025".

4.1 Panoramica dei visual e grafici disponibili in Power BI

Power BI mette a disposizione una vasta gamma di visual e grafici per rappresentare dati di vario tipo o in base alle necessità dell'analisi. Oltre ai grafici già presenti, altri possono essere importati da Power BI Visuals Marketplace. I principali grafici disponibili sono:

- *Grafico a colonne*: strumento visivo che usa barre verticali per confrontare dati tra diverse categorie. L'altezza di ogni colonna mostra il valore numerico corrispondente. Alcune varianti includono il grafico a colonne impilate o raggruppate, utile per eseguire confronti multipli.
- *Grafico a barre*: strumento visivo simile al grafico a colonne, solo che utilizza barre orizzontali; anch'esso supporta versioni impilate o raggruppate.
- *Grafico a torta*: strumento visivo a forma di cerchio diviso in fette. Ogni fetta mostra una parte rispetto a un totale del cento per cento. Esso si usa per capire il peso di singoli gruppi o delle singole categorie.
- *Grafico ad anello*: strumento visivo simile al grafico a torta, solo con una struttura ad anello.
- *Grafico a linee*: strumento visivo che utilizza dati continui come punti collegati da linee, il che lo rende ideale per la visualizzazione delle tendenze nel tempo. Può essere combinato con grafici a colonne per creare rappresentazioni miste.
- *Grafico ad area*: strumento visivo basato sul grafico a linee, con l'area tra asse e linea riempita con colori per indicare il volume. I grafici ad area evidenziano la grandezza del cambiamento nel tempo e attirano l'attenzione sul valore totale in una tendenza.

- *Istogramma*: strumento visivo che rappresenta la distribuzione di dati continui suddivisi in intervalli. Esso è spesso utilizzato per analizzare la frequenza di valori in classi specifiche.
- *Grafico a dispersione*: strumento visivo che mostra come i valori numerici lungo i due assi siano correlati. Quando i dati si intersecano sui due assi, Power BI visualizza un punto dati. Analizzando i punti dati, è possibile identificare le relazioni tra i campi sugli assi.
- *Grafico ad imbuto*: strumento visivo che consente di mostrare un processo lineare caratterizzato da fasi connesse in sequenza. Ogni fase dell'imbuto rappresenta una percentuale del totale.
- *Grafico a mappa*: strumento visivo che consente di rappresentare dati geografici e spaziali. I principali tipi sono:
  - *Mappa base*: visualizza i punti dati come bolle su una mappa, con dimensioni delle bolle che rappresentano i valori di misura.
  - *Mappa ad albero*: mostra dati gerarchici in un set di rettangoli annidati proporzionali ai valori della misura utilizzata.
  - *Mappa colorata*: usa l'ombreggiatura, l'intensità del colore o i modelli per mostrare come i valori differiscono tra le regioni geografiche.
  - *Mappa di Azure*: supporta bolle, colonne 3D, mappa termica, riempimento, marcatore, livello percorso e livelli di riferimento.
- *Grafico a matrice*: strumento visivo che si utilizza per mostrare e analizzare i dati su più dimensioni. Analogamente a una tabella pivot in Excel, la matrice consente di organizzare i dati in righe e colonne, semplificando la comprensione dei set di dati complessi.
- *Tabella*: strumento visivo contenente dati correlati in una serie logica di righe e colonne. Una tabella può contenere anche intestazioni e una riga per i totali. Le tabelle sono particolarmente utili per i confronti quantitativi in cui si analizzano molti valori per una singola categoria.
- *KPI (Key Performance Indicator)*: strumento visivo che comunica lo stato di avanzamento verso un obiettivo misurabile.
- *Card e Multi-row Card*: strumento visivo che mostra le metriche chiave in un formato flessibile. Ogni scheda rappresenta un valore di misura con immagini esplicative opzionali, etichette di riferimento e dettagli di supporto. È possibile visualizzare più schede in un singolo oggetto visivo, suddividerle in base alle categorie e personalizzare i layout in base alla progettazione del report.
- *Misuratore*: strumento visivo contraddistinto da un arco circolare per mostrare un singolo valore che misura lo stato rispetto al raggiungimento di un obiettivo o un KPI. La linea (o lancetta) del misuratore rappresenta il valore dell'obiettivo o di destinazione. L'ombreggiatura rappresenta lo stato di avanzamento rispetto al raggiungimento dell'obiettivo.
- *Albero di scomposizione*: strumento visivo che consente di visualizzare i dati attraverso più dimensioni. È una visualizzazione di Intelligenza Artificiale che aggrega automaticamente i dati e consente di approfondire le dimensioni in qualsiasi ordine.

- *Grafico a cascata*: strumento visivo che mostra un totale progressivo mentre Power BI somma e sottrae valori. Questo grafico è utile per comprendere in che modo un valore iniziale è interessato da una serie di modifiche positive e negative. Le colonne sono contraddistinte dal colore per poter notare rapidamente gli aumenti e le diminuzioni tra i dati.
- *Mappa di calore*: strumento visivo che rappresenta dati tabellari colorando le celle in funzione del valore. Esso è utile per individuare pattern e concentrazioni nei dati.
- *Grafici tramite script Python e R*: strumenti visivi che permettono l'integrazione con linguaggi di scripting, come Python e R, per creare visualizzazioni avanzate o personalizzate qualora i visual di base di Power BI non siano adatti. In questi script si possono utilizzare:
  - *Seaborn e Matplotlib (Python)*: sono librerie per la creazione di visualizzazioni statiche, animate e interattive.
  - *Pandas (Python)*: è una libreria per la manipolazione e l'analisi dei dati, fondamentale prima di richiamare altre librerie per la rappresentazione grafica.
  - *ggplot2 (R)*: è il package di riferimento del linguaggio R, per la realizzazione di grafici complessi e altamente personalizzabili.
  - *R base*: ambiente nativo di R per visualizzazioni rapide, con comandi semplici utili per un'analisi statistica classica.

4.2 Misure DAX implementate

In questa analisi sono state implementate diverse misure DAX (Data Analysis Expression) allo scopo di analizzare e visualizzare in modo chiaro le informazioni all'interno delle dashboard. Esse possono essere divise in due categorie: le misure, utilizzate per calcolare valori dinamici nei grafici e negli indicatori, e le colonne calcolate, impiegate per aggiungere nuove informazioni ai dataset e per garantire la coerenza visiva e la corretta formattazione dei report.

Di seguito si descrivono le principali misure sviluppate che sono state inserite nella tabella *_Misure*.

- *% Market Penetration*: calcola la percentuale di Paesi che hanno almeno un annuncio di lavoro rispetto al totale dei Paesi tracciati nel dataset; fornisce, quindi, la percentuale di Paesi che hanno un mercato dell'AI "attivo" (Listato 4.1).

```
1  % Market Penetration =
2  VAR _Active = CALCULATE(COUNTROWS('AI_index_db'), '
   AI_index_db'[Market Group] = "Active Market")
3  VAR _Total = COUNTROWS('AI_index_db')
4  RETURN
5  IF(ISBLANK(_Total), 0, DIVIDE(_Active, _Total, 0)) + 0
6
```

Listing 4.1: Misura DAX di *% Market Penetration*

- *Average Salary*: misura la media degli stipendi offerti da tutti gli annunci di lavoro, fornendo una stima dello stipendio per i lavoratori in ambito AI (Listato 4.2).

```
1      Average Salary = AVERAGE(ai_job_dataset1[Salary USD])
2
```

Listing 4.2: Misura DAX di *Average Salary*

- *Average Years Experience*: misura la media di anni di esperienza richiesti di tutti gli annunci di lavoro, offrendo un'indicazione del livello di esperienza complessivamente richiesto dal mercato (Listato 4.3).

```
1      Average Years Experience = AVERAGE(ai_job_dataset1[
       Years Experience])
2
```

Listing 4.3: Misura DAX di *Average Years Experience*

- *Counting of Skills*: calcola il numero totale di skill richieste dagli annunci di lavoro, fornendo un'indicazione sulle competenze più ricercate dal mercato (Listato 4.4).

```
1      Counting of Skills = COUNT(Skill_Analisis[Required
       Skills])
2
```

Listing 4.4: Misura DAX di *Counting of Skills*

- *Full-Remote Percentage*: calcola la percentuale di annunci di lavoro che possono essere svolti interamente da remoto, ovvero quando il *Remote Ratio* è pari a 100 (Listato 4.5).

```
1      Full-Remote Percentage =
2      DIVIDE(
3      CALCULATE([Total Job Postings], ai_job_dataset1[Remote
       Ratio] = 100),
4      [Total Job Postings], 0 )
5
```

Listing 4.5: Misura DAX di *Full-Remote Percentage*

- *Global Average Total Score*: calcola il valore medio del *Total Score* relativo ai soli Paesi presenti nel dataset degli annunci di lavoro, rimuovendo i filtri dinamici del report per fornire un benchmark globale di confronto sullo sviluppo dell'AI (Listato 4.6).

```
1      Global Average Total Score =
2      VAR ListOfCompanyLocation = CALCULATETABLE(VALUES('
       ai_job_dataset1'[Company Location]), REMOVEFILTERS())
3      RETURN
4      CALCULATE(
5      AVERAGE('AI_index_db'[Total score]),
6      REMOVEFILTERS(),
7      TREATAS(ListOfCompanyLocation, 'AI_index_db'[Country])
8      )
9
```

Listing 4.6: Misura DAX di *Global Average Total Score*

- *Global Benchmark*: calcola il valore medio degli stipendi, rimuovendo i filtri dinamici del report in modo da avere un benchmark globale di confronto per gli stipendi (Listato 4.7).

```
Global Benchmark =
CALCULATE(
[Average Salary],
REMOVEFILTERS('ai_job_dataset1')
)

```

Listing 4.7: Misura DAX di *Global Benchmark*

- *Hiring Velocity*: calcola la media dei giorni che intercorrono tra la pubblicazione dell'annuncio e la chiusura della posizione; è utile per valutare la reattività del mercato quando viene pubblicato un annuncio di lavoro (Listato 4.8).

```
Hiring Velocity = AVERAGE(ai_job_dataset1[Days To Apply
])

```

Listing 4.8: Misura DAX di *Hiring Velocity*

- *Key Strength Display*: cerca l'indicatore con il punteggio massimo per il Paese selezionato nel report; qualora nessun Paese sia selezionato, restituisce un messaggio di avviso che invita l'utente a effettuare una selezione (Listato 4.9).

```
Key Strength Display =
IF(
ISFILTERED('AI_index_db'[Country]),
CALCULATE(
    SELECTEDVALUE('AI_index_UNIPIVOT'[Index]),
    TOPN(
        1,
        ALL(AI_index_UNIPIVOT[Index]),
        CALCULATE(AVERAGE(AI_index_UNIPIVOT[Value]))
    )
),
"Select a Country"
)

```

Listing 4.9: Misura DAX di *Key Strength Display*

- *New Job Openings (30d)*: calcola il numero degli annunci di lavoro pubblicati nei trenta giorni prima dell'ultima data disponibile nel primo dataset. La misura permette di valutare le tendenze di breve periodo del settore AI, evidenziando eventuali fasi di espansione o contrazione del mercato (Listato 4.10).

```
New Job Openings (30d) =
VAR UltimaData = MAX(ai_job_dataset1[Posting Date])
RETURN
CALCULATE(
[Total Job Postings],
ai_job_dataset1[Posting Date] > UltimaData - 30
)

```

Listing 4.10: Misura DAX di *New Job Openings (30d)*

- *Normalized Salary*: riproporziona lo stipendio medio su una scala normalizzata da 0 a 100. Tale misura rende il dato direttamente confrontabile con gli indici presenti nel secondo dataset (Listato 4.11).

```
1  Normalized Salary =
2  VAR MinSalary = MINX(ALL('ai_job_dataset1'), '
   ai_job_dataset1'[Salary USD])
3  VAR MaxSalary = MAXX(ALL('ai_job_dataset1'), '
   ai_job_dataset1'[Salary USD])
4  VAR CurrentSalary = [Average Salary]
5  RETURN
6  DIVIDE(
7  CurrentSalary - MinSalary,
8  MaxSalary - MinSalary
9  ) * 100
10
```

Listing 4.11: Misura DAX di *Normalized Salary*

- *Overall Score Display*: calcola la media del *Total Score* del Paese selezionato, se non viene selezionato alcun Paese restituisce la stringa "Select a Country" (Listato 4.12).

```
1  Overall Score Display =
2  IF(
3  ISFILTERED('AI_index_db'[Country]),
4  FORMAT(AVERAGE('AI_index_db'[Total Score]), "0.0"),
5  "Select a Country"
6  )
7
```

Listing 4.12: Misura DAX di *Overall Score Display*

- *Peak Salary*: individua il salario massimo registrato all'interno del primo dataset (Listato 4.13).

```
1  Peak Salary = MAX(ai_job_dataset1[Salary USD])
2
```

Listing 4.13: Misura DAX di *Peak Salary*

- *Remote Adoption Rate*: calcola la percentuale degli annunci di lavoro che possono essere svolti in modalità ibrida o interamente da remoto (*Remote Ratio* pari a 50 o 100) rispetto al totale delle offerte rilevate (Listato 4.14).

```
1  Remote Adoption Rate =
2  DIVIDE(
3  CALCULATE([Total Job Postings], ai_job_dataset1[Remote
   Ratio] > 0),
4  [Total Job Postings],
5  0
6  )
7
```

Listing 4.14: Misura DAX di *Remote Adoption Rate*

- *Salary Gap*: calcola la differenza tra lo stipendio medio e il *Global Benchmark*. La misura evidenzia se il segmento analizzato si posiziona al di sopra o al di sotto del parametro di riferimento (Listato 4.15).

```
Salary Gap =
VAR MediaCorrente = AVERAGE('ai_job_dataset1'[Salary
USD])
VAR Differenza = MediaCorrente - [Global Benchmark]
RETURN
Differenza
```

Listing 4.15: Misura DAX di *Salary Gap*

- *Strategic Gap*: calcola la differenza tra la media del *Government Strategy* per i Paesi caratterizzati da un mercato attivo rispetto a quelli con mercato dormiente. Trattandosi di una misura statica, non viene influenzata dai filtri applicati al report, fornendo un riferimento globale costante (Listato 4.16).

```
Strategic Gap =
VAR _AvgActiveGlobal =
CALCULATE(
    AVERAGE('AI_index_db'[Government Strategy]),
    'AI_index_db'[Market Group] = "Active Market",
    ALL('AI_index_db')
)
VAR _AvgDormantGlobal =
CALCULATE(
    AVERAGE('AI_index_db'[Government Strategy]),
    'AI_index_db'[Market Group] = "Dormant Market",
    ALL('AI_index_db')
)
RETURN
_AvgActiveGlobal - _AvgDormantGlobal
```

Listing 4.16: Misura DAX di *Strategic Gap*

- *Target Benchmark*: calcola il valore massimo dell'indice considerando l'intero insieme di Paesi, ma mantenendo attivi i filtri selezionati sui cluster. La misura definisce il traguardo di riferimento all'interno del gruppo di analisi considerato (Listato 4.17).

```
Target Benchmark =
CALCULATE(
MAX('AI_index_UNIPIVOT'[Value]),
ALL('AI_index_db'[Country]),
ALLSELECTED('AI_index_db'[Cluster])
)
```

Listing 4.17: Misura DAX di *Target Benchmark*

- *Top Hiring Country*: restituisce il nome del Paese che presenta il maggior numero di annunci di lavoro (Listato 4.18).

```
Top Hiring Country =
CALCULATE(
SELECTEDVALUE(ai_job_dataset1[Company Location],"No
Data"),
TOPN(
    1,
```

```
6           VALUES(ai_job_dataset1[Company Location]),
7           CALCULATE([Total Job Postings]),
8           DESC
9       )
10      )
11
```

Listing 4.18: Misura DAX di *Top Hiring Country*

- *Total Job Postings*: calcola il numero complessivo di annunci di lavoro presenti nel primo dataset (Listato 4.19).

```
1       Total Job Postings = COUNTROWS(ai_job_dataset1)
2
```

Listing 4.19: Misura DAX di *Total Job Postings*

- *Weakest Pillar (Dormant)*: restituisce il nome del pilastro strategico che registra il valore medio più basso tra i Paesi con un mercato dormiente (Listato 4.20).

```
1       Weakest Pillar (Dormant) =
2       VAR _Talent = CALCULATE(AVERAGE('AI_index_db'[Talent]),
        'AI_index_db'[Market Group] = "Dormant Market")
3       VAR _Infras = CALCULATE(AVERAGE('AI_index_db'[
        Infrastructure]), 'AI_index_db'[Market Group] = "Dormant
         Market")
4       VAR _Research = CALCULATE(AVERAGE('AI_index_db'[
        Research]), 'AI_index_db'[Market Group] = "Dormant
        Market")
5       VAR _Comm = CALCULATE(AVERAGE('AI_index_db'[Commercial
        ]), 'AI_index_db'[Market Group] = "Dormant Market")
6       VAR _GovStrategy = CALCULATE(AVERAGE('AI_index_db'[
        Government Strategy]), 'AI_index_db'[Market Group] = "
        Dormant Market")
7       VAR _Dev= CALCULATE(AVERAGE('AI_index_db'[Development])
        , 'AI_index_db'[Market Group] = "Dormant Market")
8       VAR _OpEnv= CALCULATE(AVERAGE('AI_index_db'[Operating
        Environment]), 'AI_index_db'[Market Group] = "Dormant
        Market")
9       VAR _MinValue = MINX({_Talent, _Infras, _Research,
        _Comm,_GovStrategy,_Dev,_OpEnv}, [Value])
10      RETURN
11      SWITCH( TRUE(),
12      _MinValue = _Talent, "Talent",
13      _MinValue = _Infras, "Infrastructure",
14      _MinValue = _Research, "Research",
15      _MinValue = _Comm, "Commercial",
16      _MinValue = _GovStrategy, "Government Strategy",
17      _MinValue = _Dev, "Development",
18      _MinValue = _OpEnv, "Operating Environment",
19      "Balanced")
20
```

Listing 4.20: Misura DAX di *Weakest Pillar (Dormant)*

- *World Rank Display*: restituisce la posizione in graduatoria del Paese selezionato in base al valore del suo *Total Score*; qualora non sia selezionato alcun Paese, mostra il messaggio "Select a Country" (Listato 4.21).

```
World Rank Display =
IF(
ISFILTERED('AI_index_db'[Country]),
"Rank: " & RANKX(ALL('AI_index_db'), CALCULATE(AVERAGE
('AI_index_db'[Total Score])),, DESC),
"Select a Country"
)

```

Listing 4.21: Misura DAX di *World Rank Display*

Di seguito si descrivono le principali colonne calcolate.

- *Market Group*: colonna inserita nel secondo dataset per categorizzare ciascun Paese come mercato attivo (*Active Market*) o dormiente (*Dormant Market*) in funzione della presenza o assenza di offerte di lavoro registrate (Listato 4.22).

```
Market Group =
VAR _NumJobs = COUNTROWS(RELATEDTABLE('ai_job_dataset1
'))
RETURN
IF(_NumJobs > 0 || NOT(ISBLANK(_NumJobs)), "Active
Market", "Dormant Market")

```

Listing 4.22: Colonna calcolata *Market Group*

- *Month_EN*: colonna calcolata definita per convertire la data di pubblicazione dell'annuncio (*Posting Date*) nel nome del mese corrispondente in lingua inglese (*en-US*). L'integrazione di questa colonna assicura l'omogeneità linguistica di tutte le componenti grafiche della dashboard, evitando traduzioni automatiche basate sulla lingua preimpostata di Power BI (Listato 4.23).

```
Month_EN = FORMAT('ai_job_dataset1'[Posting Date], "
MMMM", "en-US")

```

Listing 4.23: Colonna calcolata *Month_EN*

- *Sort Month Date*: colonna calcolata che estrae l'indice numerico del mese (1-12) dalla data di pubblicazione dell'annuncio. Viene utilizzata come criterio di ordinamento per la colonna *Month_EN*, evitando l'ordinamento alfabetico nei grafici e garantendo la corretta sequenza dei mesi (Listato 4.24).

```
Sort Month Date = MONTH('ai_job_dataset1'[Posting Date
])

```

Listing 4.24: Colonna calcolata *Sort Month Date*

- *Year_Filter*: colonna calcolata definita per estrarre il valore numerico dell'anno dalla data di pubblicazione dell'annuncio (*Posting Date*); è utile per eseguire dei filtraggi temporali anche in contemporanea con la colonna *Month_EN* (Listato 4.25).

```
Year_Filter = YEAR('ai_job_dataset1'[Posting Date])

```

Listing 4.25: Colonna calcolata *Year_Filter*

4.3 Descrizione e analisi delle dashboard realizzate

In questa sezione vengono presentate le prime quattro dashboard interattive con lo scopo di esplorare e analizzare i dati raccolti nei due dataset, opportunamente ripuliti e trasformati.

Le dashboard sono state realizzate a partire dai dati e dalle misure DAX ottenute dal dataset *Global AI Job Market and Salary Trends 2025*. Questo dataset raccoglie diversi annunci di lavoro tra il 2024 e i primi mesi del 2025 permettendo, quindi, un'analisi microscopica sul mercato occupazionale in ambito AI.

Le dashboard sono:

- *Dashboard 1 - Global Market and Remote Landscape*: analizza il mercato del lavoro nel settore dell'Intelligenza Artificiale e l'impatto del lavoro da remoto. Il suo obiettivo è misurare il numero degli annunci di lavoro e la loro diffusione da remoto, mappando la distribuzione geografica delle offerte. Inoltre, individua i ruoli più richiesti dalle aziende, mette a confronto le fasce salariali in base alla modalità lavorativa (in presenza, ibrida o remota) e mostra come si distribuiscono le assunzioni tra piccole, medie e grandi imprese.

- *Dashboard 2 - Skills and Compensation Analysis*: si concentra sul legame tra le skill più richieste nel campo dell'Intelligenza Artificiale e il loro valore economico sul mercato. L'obiettivo è individuare le competenze più importanti per i professionisti del settore, valutando al contempo l'impatto che l'esperienza, il livello di istruzione e il settore industriale hanno sugli stipendi.

- *Dashboard 3 - Skills and Educational Mapping*: descrive il ruolo del percorso di studi nell'accesso al mercato dell'AI e nella successiva crescita professionale. L'obiettivo è valutare quanto il titolo accademico incida sulle opportunità di carriera e sulle richieste delle aziende di varie dimensioni, analizzando anche la relazione tra livello di istruzione, stipendi, benefit aziendali e il tipo di industria.

- *Dashboard 4 - Market Pulse and Hiring Dynamic*: mostra i tempi di reclutamento e il ritmo delle assunzioni nel settore dell'Intelligenza Artificiale. L'obiettivo è misurare la velocità con cui le aziende selezionano il personale e la frequenza dei nuovi annunci, valutando come variano le retribuzioni in base al ruolo, al tipo di contratto e alla crescita dell'esperienza nel tempo.

In generale, queste quattro dashboard rappresentano un punto di partenza fondamentale per ottenere una comprensione ad alto livello del mercato dell'Intelligenza Artificiale nel periodo considerato. I KPI e i grafici selezionati permettono di sintetizzare le informazioni principali in modo chiaro e immediato, supportando l'analisi dei dati e consentendo di individuare le dinamiche chiave del settore.

4.3.1 Dashboard 1: Global Market and Remote Landscape

Questa dashboard ha l'obiettivo di offrire una visione d'insieme del mercato del lavoro nel campo dell'Intelligenza Artificiale. Lo scopo è analizzare la diffusione delle offerte a livello globale, il peso del lavoro da remoto e le differenze di stipendio tra le varie realtà aziendali.

Inoltre, la dashboard permette di analizzare se la dimensione dell'azienda incide sulla ricerca di questi specialisti e se il lavoro da remoto comporti o meno una penalizzazione sullo stipendio rispetto alle posizioni in presenza o ibride.

La dashboard in oggetto è mostrata in Figura 4.1.

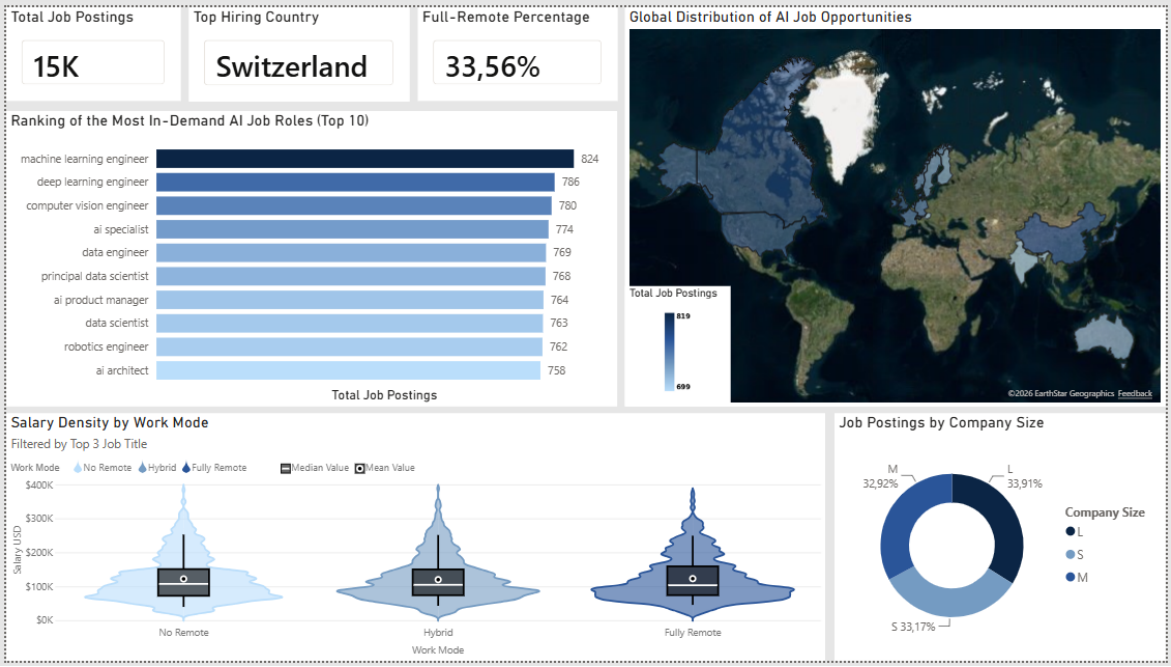


Figura 4.1: Global Market and Remote Landscape

Key Performance Indicator (KPI) utilizzati

- *Total Job Postings*: 15.000 annunci di lavoro.
- *Top Hiring Country*: la Svizzera risulta il Paese che ha pubblicato più annunci di lavoro.
- *Full-Remote Percentage*: il 33.56% degli annunci consente di svolgere il lavoro da remoto.

Grafici utilizzati

- *Mappa colorata*: mostra la distribuzione dei Paesi in base al numero di annunci pubblicati; come indicato dalla legenda, una tonalità più scura corrisponde a un maggior numero di offerte, mentre le aree non colorate evidenziano l'assenza di dati nel dataset considerato.
- *Grafico a barre orizzontali*: rappresenta i primi dieci titoli lavorativi più ricercati, disposti in ordine decrescente in base al numero di annunci presenti nel dataset.
- *Violin Plot*: mostra la distribuzione degli stipendi per le differenti modalità lavorative sui primi tre ruoli del dataset, permettendo di analizzare la densità dei dati e i valori statistici, come media e mediana.
- *Grafico ad anello*: illustra la percentuale di annunci pubblicati dalle aziende in base alla loro dimensione (piccole, medie, grandi).

Conclusioni sull'analisi della Dashboard 1

L'analisi svolta nella prima dashboard permette di tracciare un quadro generale sul mercato del lavoro nell'ambito dell'AI tra l'inizio del 2024 e i primi mesi del 2025.

Dai KPI principali si nota una buona richiesta di figure specializzate in questo settore, pur considerando che il dataset copre un numero limitato di Paesi. Tra le nazioni analizzate, la Svizzera risulta al primo posto per numero di annunci pubblicati, seguita da Stati Uniti e Canada.

Per quanto riguarda l'organizzazione del lavoro, i dati mostrano che circa un terzo delle offerte prevede la modalità da remoto. Osservando il Violin Plot, si nota anche che il lavoro a distanza non porta a una riduzione degli stipendi rispetto al lavoro in presenza. Inoltre, l'analisi per dimensione aziendale evidenzia come la ricerca di profili AI sia diffusa in modo omogeneo, sia nelle piccole e medie imprese che in quelle di grandi dimensioni.

Infine, analizzando le singole professioni, le figure più richieste sono il Machine Learning Engineer, il Deep Learning Engineer e il Computer Vision Engineer.

4.3.2 Dashboard 2: Skills and Compensation Analysis

Questa dashboard è stata progettata per evidenziare le competenze tecniche maggiormente richieste nel settore dell'Intelligenza Artificiale e per valutare il loro valore economico sul mercato. Attraverso i KPI e i grafici integrati, è possibile comprendere rapidamente come l'esperienza, il livello di istruzione e il settore industriale influenzino gli stipendi offerti.

La dashboard in oggetto è riportata in Figura 4.2.

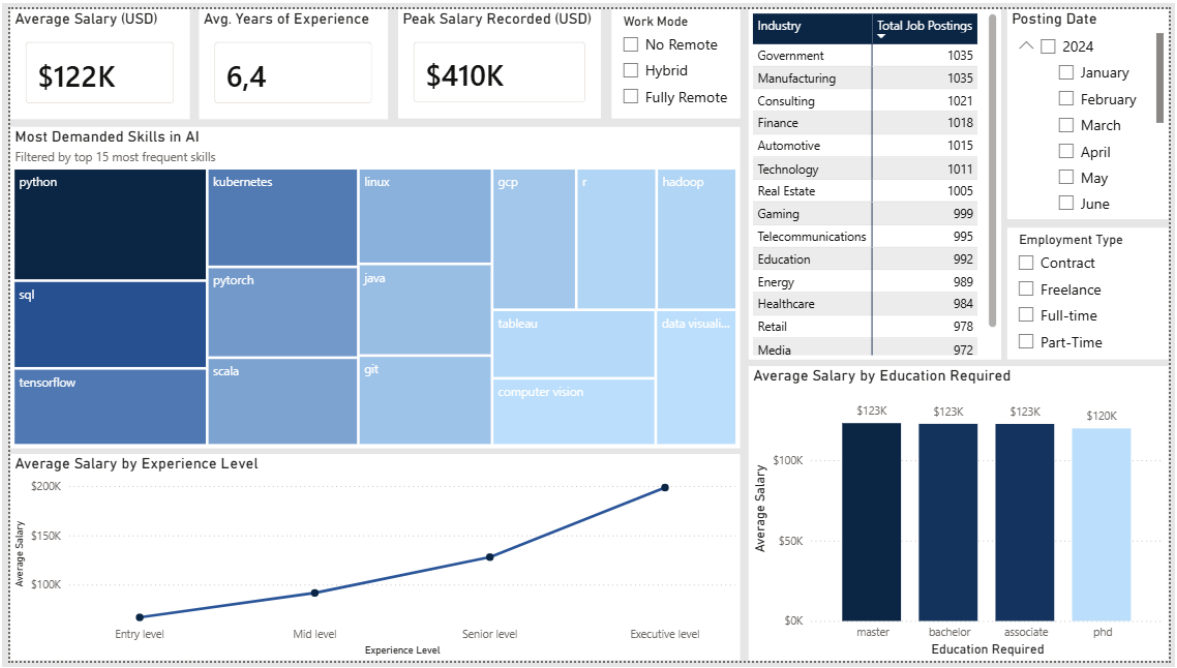


Figura 4.2: Skills and Compensation Analysis

Key Performance Indicator (KPI) utilizzati

- *Average Salary (USD)*: $122.000.
- *Avg. Years of Experience*: 6,4 anni di esperienza.
- *Peak Salary Recorded (USD)*: $410.000.

Grafici utilizzati

- *Mappa ad albero*: evidenzia le quindici skill maggiormente richieste sul mercato, dove l'ampiezza e il colore dei riquadri riflettono la frequenza di ciascuna competenza.
- *Grafico a linee*: mostra l'andamento dello stipendio medio in funzione del livello di esperienza, a partire dall'Entry level fino all'Executive level.

- *Tabella*: mostra il numero totale di annunci di lavoro per ciascun settore industriale, consentendo di individuare gli ambiti con la maggiore possibilità di assunzione.
- *Grafico a barre verticali*: confronta lo stipendio medio in base al titolo di studio richiesto, evidenziando le differenze tra master, bachelor, associate e phd.

Conclusioni sull'analisi della Dashboard 2

L'analisi svolta nella seconda dashboard mette in luce una forte differenza tra lo stipendio medio e i valori di picco registrati sul dataset. I KPI evidenziano inoltre come le aziende ricerchino in media personale che abbia almeno sei anni di esperienza. Osservando la mappa ad albero, le skill che registrano la frequenza più elevata negli annunci risultano essere Python, SQL e TensorFlow, che si confermano strumenti centrali per chi opera nel settore.

Per quanto riguarda l'andamento degli stipendi, il grafico a linee mostra una crescita costante del salario medio nei primi livelli di esperienza (Entry, Mid e Senior level), per poi registrare un aumento più netto quando si passa alle posizioni di livello Executive.

La tabella evidenzia come la maggiore concentrazione di offerte di lavoro risieda nel settore governativo, del manifatturiero e della consulenza.

Infine, dall'analisi del grafico a barre emerge uno stipendio medio fondamentalmente allineato rispetto al titolo di studio richiesto, con una minima diminuzione per le posizioni che richiedono un PhD. Va tuttavia notato come il dato relativo al PhD possa risentire sia del minor numero di annunci presenti nel dataset per questo livello di studio, sia del fatto che tali ruoli sono spesso legati ad ambiti di ricerca o al settore pubblico, dove gli stipendi medi tendono a essere più contenuti rispetto alle posizioni gestionali del privato.

In sintesi, i dati mostrano un settore in cui il mercato valorizza soprattutto l'esperienza pratica e la padronanza degli strumenti tecnici, mentre il titolo accademico incide in misura minore sui livelli retributivi di partenza.

4.3.3 Dashboard 3: Skills and Educational Mapping

La terza dashboard approfondisce la relazione tra la formazione accademica e la crescita professionale all'interno del mercato dell'Intelligenza Artificiale. Essa permette di analizzare come il livello di istruzione influenzi l'evoluzione della carriera, le preferenze di assunzione in base alla dimensione aziendale, la distribuzione degli stipendi e dei benefit nei vari settori.

La dashboard in oggetto viene mostrata in Figura 4.3.

Grafici utilizzati

- *Grafico a nastri*: mostra il flusso delle posizioni lavorative per ciascun livello di studio, evidenziando come varia la distribuzione dell'esperienza al variare del titolo accademico.
- *Grafico a dispersione*: mette in relazione il punteggio dei benefit con lo stipendio medio delle varie posizioni lavorative, distinguendo i punti in base al titolo di studio richiesto.
- *Grafico a barre in pila*: mostra come si distribuiscono le richieste dei vari titoli di studio tra piccole, medie e grandi aziende, mettendo in evidenza a quali qualifiche sono più interessate le diverse realtà aziendali
- *Mappa di calore*: analizza lo stipendio medio nei vari settori industriali in base al titolo di studio, mettendo in risalto le combinazioni con le retribuzioni più alte.

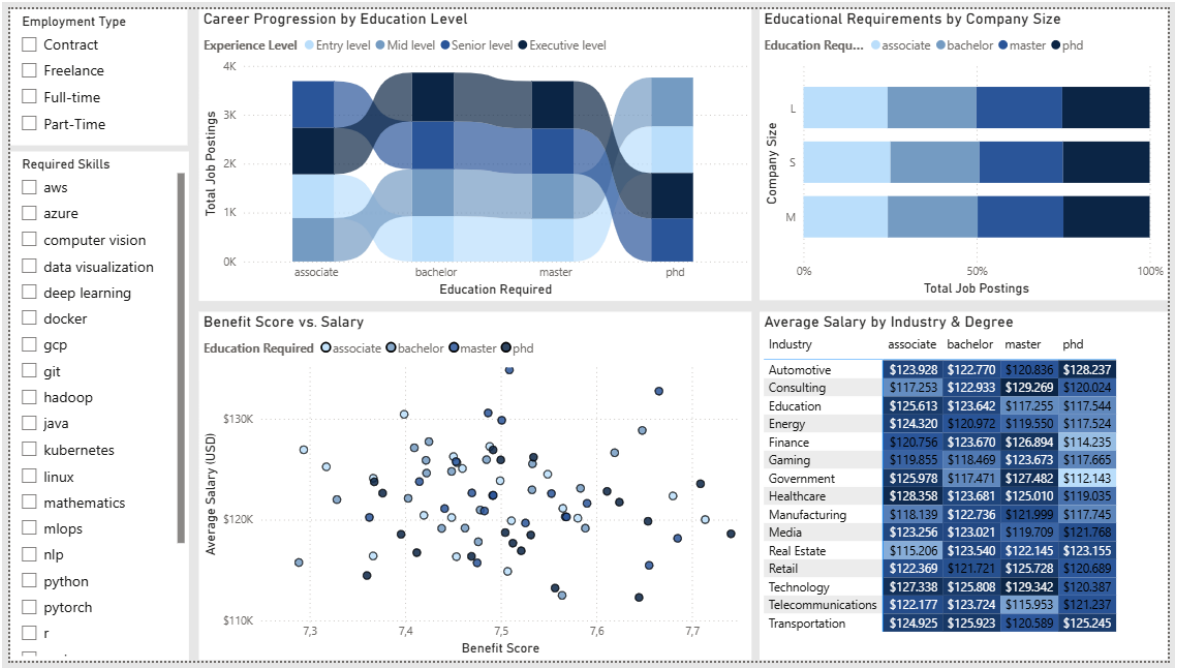


Figura 4.3: Skills and Educational Mapping

Conclusioni sull'analisi della Dashboard 3

L'analisi svolta nella terza dashboard si concentra su come la formazione possa influenzare gli stipendi e le opportunità di carriera. Infatti, nel grafico a nastri si nota come i candidati con Bachelor e Master occupino principalmente posizioni Senior ed Executive, mentre i PhD risultano più concentrati in ruoli specialistici a inizio o metà carriera; si osserva, inoltre, che il numero di annunci che richiedono questo titolo è inferiore rispetto agli altri.

Il grafico a dispersione mostra che il punteggio dei benefit si attesta prevalentemente in una fascia media (tra 7,3 e 7,7) senza evidenziare un legame diretto con lo stipendio, dato che posizioni con stipendi elevati si riscontrano sia a livelli di benefit contenuti che di benefit più alti. A influenzare il benefit sembrano essere, quindi, il livello di istruzione e il ruolo lavorativo: il valore di benefit più alto si è registrato, infatti, per un NLP Engineer con PhD.

Dal grafico a barre in pila si osserva che la richiesta di titoli di studio è uniforme indipendentemente dalla dimensione dell'impresa; piccole, medie e grandi aziende mostrano percentuali di assunzione quasi identiche tra i vari livelli di istruzione, anche se ciò può essere dovuto al campione di dati raccolto nel dataset.

Infine, dalla mappa di calore si nota che la retribuzione media si mantiene stabile tra i $115.000 e i $129.000 su quasi tutti i settori, con i picchi più alti registrati nell'Automotive per i PhD e nei settori del consulting e della technology per i Master.

4.3.4 Dashboard 4: Market Pulse and Hiring Dynamic

Questa dashboard è progettata per analizzare l'efficienza dei processi di selezione e le dinamiche degli stipendi nel settore dell'Intelligenza Artificiale. Le visualizzazioni proposte offrono una panoramica sui tempi medi di assunzione, sulla frequenza delle nuove offerte di lavoro e sulle correlazioni tra retribuzione, tipologia contrattuale ed esperienza dei candidati.

La dashboard in oggetto è riportata in Figura 4.4.

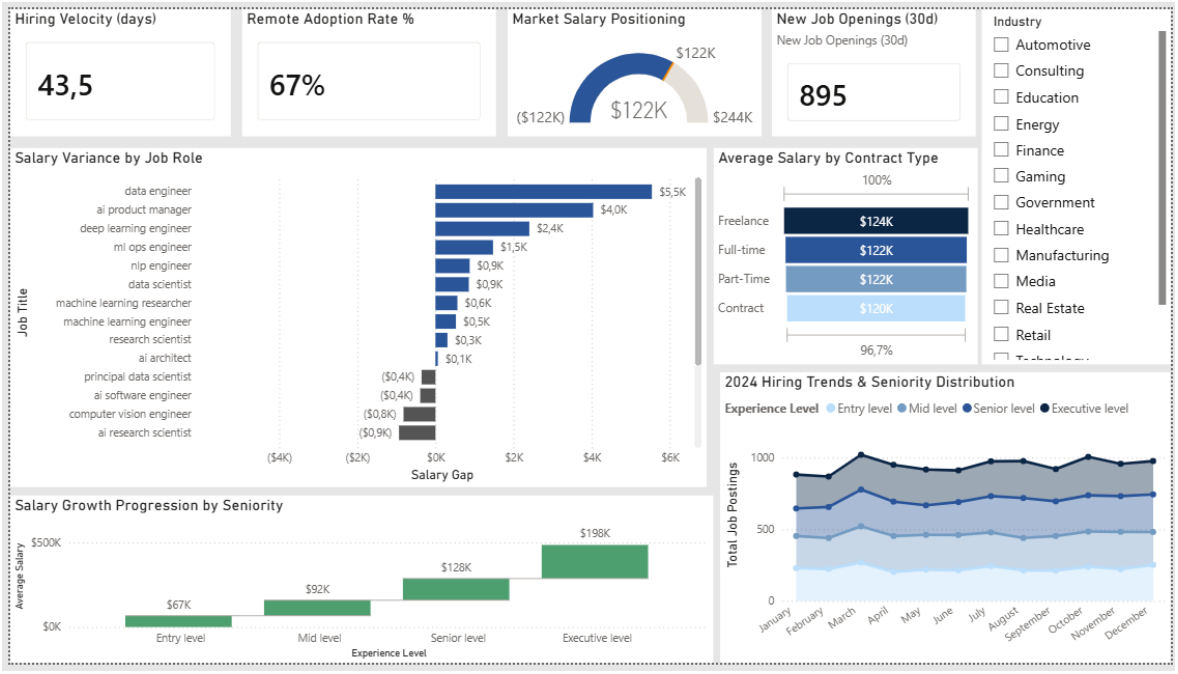


Figura 4.4: Market Pulse and Hiring Dynamic

Key Performance Indicator (KPI) utilizzati

- *Hiring Velocity (days)*: 43,5 giorni trascorsi tra la pubblicazione dell'annuncio e l'assunzione.
- *Remote Adoption Rate%*: 67%.
- *New Job Openings (30d)*: 895 annunci di lavoro sono stati pubblicati negli ultimi trenta giorni raccolti nel dataset.

Grafici utilizzati

- *Misuratore*: mostra la collocazione dello stipendio medio rispetto al target di mercato. In assenza di filtri applicati, il valore medio rimane fisso sul target.
- *Grafico a barre orizzontali*: mostra la differenza degli stipendi rispetto al valore medio globale per ciascun ruolo lavorativo. Le barre verso sinistra indicano i ruoli con gap negativo (stipendi sotto la media), mentre quelle a destra evidenziano i ruoli con valori positivi.
- *Grafico a cascata*: analizza l'incremento dello stipendio in base al livello di esperienza. Ogni gradino rappresenta l'aumento di retribuzione da un livello al successivo, evidenziando come l'anzianità incide sui salari.
- *Grafico a imbuto*: illustra la distribuzione degli stipendi medi tra le diverse tipologie di contratto (Full-time, Part-time, Freelance, Contract). Permette di confrontare stipendi in base al tipo di contratto lavorativo.
- *Grafico ad area*: mostra il totale mensile degli annunci di lavoro nel 2024 divisi per livello di esperienza. Permette di valutare come cambia la richiesta nel corso dell'anno per i diversi profili.

Conclusioni sull'analisi della Dashboard 4

L'analisi svolta nella quarta dashboard si concentra sulle dinamiche di assunzione e sui livelli salariali del settore. I KPI mostrano un tempo medio di assunzione di 43,5 giorni, con una quota di lavoro da remoto o ibrido del 67% e 895 annunci pubblicati negli ultimi trenta giorni, a conferma di un mercato flessibile e con tempistiche di selezione ben definite.

Filtrando i dati per i diversi settori, il misuratore mostra la variazione dello stipendio medio rispetto al valore globale. Questa differenza viene approfondita nel grafico a barre orizzontali per ogni ruolo lavorativo: l'aumento maggiore rispetto alla media si registra per il Data Engineer (+$5.500), mentre il divario negativo più accentuato riguarda il Data Analyst (-$5.000).

Il grafico a cascata mostra la crescita degli stipendi all'aumentare dell'esperienza, partendo da $67.000 per i profili Entry level fino ad arrivare a $198.000 per i ruoli Executive. Dal grafico sulle tipologie contrattuali si nota, invece, che le retribuzioni medie variano poco tra le diverse forme lavorative (dai $120.000 dei Contract ai $124.000 dei Freelance), dimostrando che lo stipendio dipende più dal livello di esperienza che dal tipo di contratto.

Infine, dal grafico ad area si osserva che il numero di annunci si è mantenuto stabile per tutto il 2024, con la maggior parte delle offerte rivolta a profili di livello Mid e Senior.

CAPITOLO 5

Progettazione e realizzazione delle dashboard complesse

In questo capitolo vengono presentate la progettazione e l'implementazione delle quattro dashboard complesse realizzate nell'ambito dello studio. Dopo un'introduzione ai visual e ai grafici avanzati impiegati, si passa all'analisi dettagliata di ciascuna dashboard. A differenza delle precedenti, basate esclusivamente sul dataset "Global AI Job Market and Salary Trends 2025", queste dashboard integrano e mettono in relazione le informazioni di entrambi i dataset considerati o si focalizzano sui dati dell'AI Global Index.

5.1 Visual e grafici avanzati

Power BI, oltre a mettere a disposizione degli utenti visual e grafici di base, permette di estenderne le funzionalità importandone altri da Power BI Visuals Marketplace oppure sviluppandone di personalizzati tramite script Python o R.

L'impiego di grafici di base si è rivelato insufficiente per rappresentare in modo efficace relazioni multidimensionali, distribuzioni di densità e confronti complessi tra le misure DAX e i dati relativi al mercato del lavoro con gli indicatori raccolti in *AI Global Index*. Per ovviare a tali limitazioni e garantire una capacità di analisi più approfondita, sono state integrate diverse tipologie di grafici avanzati. In particolare sono stati utilizzati:

- *Bullet Chart*: consente di confrontare un indicatore con un valore obiettivo e di contestualizzarlo rispetto a diverse fasce di prestazione. Rispetto ai tradizionali misuratori, offre una resa visiva più dettagliata, rendendo immediata la valutazione del risultato ottenuto.
- *Radar Chart*: consente di rappresentare contemporaneamente tre o più variabili quantitative su assi radiali distinti che si sviluppano da un centro comune. I punti tracciati su ciascun asse vengono congiunti per formare un poligono, rendendo immediato il confronto di valori e offrendo una panoramica chiara dei punti di forza o delle criticità nelle prestazioni.
- *Sankey Diagram*: consente di individuare chiaramente le fonti, le destinazioni e le fasi intermedie, nonché il flusso di informazioni tra di esse. Risulta efficace per rappresentare relazioni complesse che coinvolgono molteplici passaggi e intermediari.
- *Bubble Matrix*: organizza i dati all'interno di una griglia di righe e colonne. Sfrutta la dimensione e il colore delle bolle nelle intersezioni per mostrare contemporaneamente

sia variabili quantitative che categoriche, consentendo di confrontare più informazioni in un unico grafico.

- *Hexbin Plot*: raggruppa i dati all'interno di una griglia di celle esagonali, associando a ciascun esagono un colore più o meno intenso a seconda della quantità di valori contenuti. Risulta molto utile per gestire il problema della sovrapposizione dei dati nei grafici a dispersione quando si lavora con grandi dataset.
- *Correlation Matrix*: organizza in una tabella o mappa di calore i coefficienti di correlazione tra più variabili numeriche del dataset. I valori variano da -1 a 1, dove 1 indica una perfetta correlazione positiva, -1 una perfetta correlazione negativa e 0 l'assenza di legame lineare; è utile per l'analisi esplorativa dei dati e per individuare rapidamente quali misure mostrano relazioni significative di esse.

5.2 Descrizione e analisi delle dashboard realizzate

In questa sezione vengono presentate le quattro dashboard complesse sviluppate per analizzare i dati raccolti nei due dataset. L'impiego di visualizzazioni avanzate ha permesso di far emergere relazioni e insight che non sarebbero stati immediatamente evidenti attraverso l'uso di grafici standard.

Le dashboard sono state realizzate utilizzando combinazioni di dati e misure DAX provenienti sia dal dataset *AI Global Index* sia dal dataset *Global AI Job Market and Salary Trends 2025* (ad eccezione della Dashboard 5, basata esclusivamente sul primo). Il dataset *AI Global Index* raccoglie l'indice globale sull'AI e altri sette indici focalizzati su innovazione, sviluppo e investimenti nei diversi Paesi. Grazie ad esso si passa da un'analisi microscopica, condotta nelle prime dashboard sul mercato del lavoro, a una prospettiva macroscopica sull'adozione dell'AI a livello nazionale, al fine di valutare come questi indici influenzino i singoli mercati.

Le dashboard sono:

- *Dashboard 5 - Strategic Pillars and Global Positioning*: si concentra sul posizionamento dei vari Paesi nello scenario globale dell'AI e sulla loro suddivisione all'interno dei diversi cluster regionali. L'obiettivo è quello di fornire un quadro chiaro delle prestazioni nazionali mediante le misure principali, mettendo in luce i divari presenti sui pilastri strategici (come infrastrutture, talenti e ricerca) e definendo le prestazioni di ciascuna nazione nei diversi ambiti d'analisi.
- *Dashboard 6 - Global Salaries and Cluster Performance*: confronta gli stipendi con i livelli di sviluppo tecnologico dei vari Paesi. L'obiettivo è quello di capire come cambiano gli stipendi per le 15 figure più richieste all'interno dei diversi cluster, analizzando al tempo stesso la concentrazione delle offerte rispetto alla maturità delle infrastrutture nazionali e il legame tra lo stipendio normalizzato e i principali indici di performance (infrastruttura, talenti e ricerca).
- *Dashboard 7 - Predictive Career and Market Growth*: approfondisce le prospettive di crescita professionale e l'impatto delle dinamiche di mercato sull'evoluzione della carriera. L'obiettivo è quello di studiare come si legano tra loro l'esperienza lavorativa, gli stipendi e il livello di maturità tecnologica dei vari Paesi, valutando le proiezioni di guadagno e confrontando l'andamento degli stipendi nelle 15 nazioni con i compensi più alti rispetto al loro effettivo sviluppo nell'AI.
- *Dashboard 8 - Market Dynamics and AI Readiness Landscape*: evidenzia il livello di preparazione tecnologica dei vari Paesi mettendo a confronto i mercati attivi con quelli

dormienti. L'obiettivo è quello di analizzare la diffusione dell'AI su scala globale e misurare il divario strategico tra le nazioni, osservando come la pianificazione dei governi incida sui diversi cluster per individuare quali pilastri penalizzino di più i contesti meno sviluppati.

5.2.1 Dashboard 5: Strategic Pillars and Global Positioning

Questa dashboard sposta l'analisi su un punto di vista globale, proponendo un confronto diretto tra i diversi ecosistemi nazionali dell'Intelligenza Artificiale. Strutturata per offrire una visione d'insieme immediata, essa permette di valutare il livello di sviluppo dei singoli Paesi e di comprendere come le scelte strategiche locali influenzino la competitività globale.

La dashboard in oggetto è riportata in Figura 5.1.

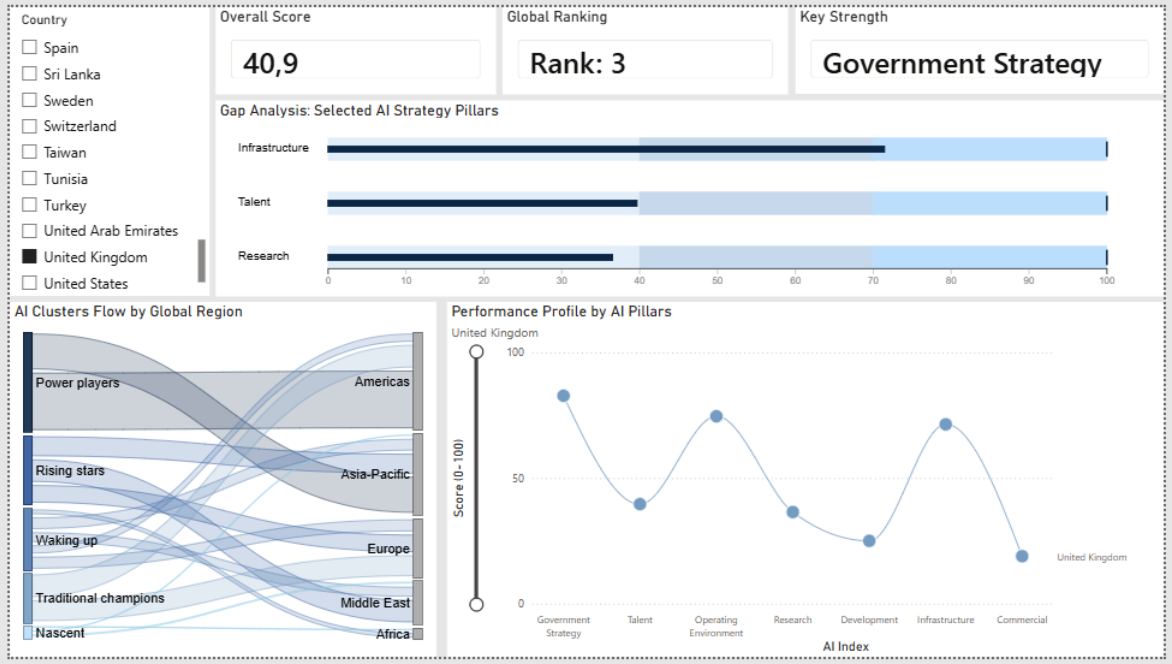


Figura 5.1: Strategic Pillars and Global Positioning

Key Performance Indicator (KPI) utilizzati

I KPI riportati in Figura 5.1 mostrano i valori calcolati applicando il filtro per Paese su United Kingdom:

- *Overall Score*: 40,9; rappresenta il valore del Total Score del Paese selezionato.
- *Global Ranking*: Rank 3; rappresenta la posizione tra i 62 Paesi presenti nel dataset in base al Total Score.
- *Key Strength*: Government Strategy; registra il valore più alto per il Paese selezionato.

Grafici utilizzati

- *Bullet Chart*: mostra le prestazioni del Paese selezionato rispetto alla nazione leader sui singoli indici di valutazione. La barra blu indica il valore reale registrato dal Paese, mentre la linea verticale segna il benchmark di riferimento, calcolato sui punteggi

ottenuti dalla nazione al primo posto. Gli indici che vengono confrontati si limitano a Talent, Infrastructure e Research, poiché sono quelli che permettono un confronto più immediato con i valori del primo dataset.

- *Sankey Diagram*: traccia la relazione tra la posizione geografica delle nazioni, i rispettivi cluster di appartenenza e il valore dell'indice globale sull'AI ottenuto. Il grafico permette di osservare l'effettiva dominanza di alcune regioni su specifici livelli di sviluppo, rendendo subito evidente se a un punteggio complessivo alto corrisponda sempre l'inserimento nei gruppi di testa o se vi siano eccezioni significative.
- *Grafico a linee*: mostra la distribuzione dei punteggi sui diversi indici di valutazione. Di base il grafico confronta le prime cinque nazioni per Total Score, mentre applicando il filtro sul singolo Paese, come in Figura 5.1, consente di analizzarne il profilo specifico lungo tutti gli indicatori.

Conclusioni sull'analisi della Dashboard 5

L'analisi condotta nella quinta dashboard si focalizza sugli indici relativi alle singole nazioni e sul loro posizionamento globale nello sviluppo dell'Intelligenza Artificiale. L'utilizzo del filtro per Paese consente di isolare le metriche di dettaglio per ciascuna realtà.

Selezionando lo United Kingdom, i KPI evidenziano che, con un Total Score di 40,9, il Paese si colloca al terzo posto nel ranking mondiale, individuando nelle strategie governative (Government Strategy) il proprio principale punto di forza. Dalla bullet chart si osserva come anche il livello infrastrutturale si attesti su una fascia alta, mentre indici chiave come il talento e la ricerca rimangono confinati nella fascia media; la forte spinta delle politiche governative emerge, quindi, come la causa primaria dell'elevato posizionamento in classifica.

Il grafico a linee mostra come i valori più bassi registrati dal Paese corrispondano ai pilastri Development e Commercial, individuando in queste aree i principali margini di miglioramento per scalare ulteriormente la graduatoria globale.

Infine, il sankey diagram evidenzia l'appartenenza dello United Kingdom al gruppo dei Power player all'interno del contesto europeo, riflettendo la struttura complessiva del dataset in cui i leader di primo piano si concentrano tra America ed Europa, i Paesi Rising stars e Waking up si distribuiscono, prevalentemente, tra Europa e Asia-Pacific, e le nazioni Nascent si collocano in prevalenza tra Africa e Medio Oriente.

5.2.2 Dashboard 6: Global Salaries and Cluster Performance

La sesta dashboard approfondisce il legame tra gli stipendi nell'ambito dell'Intelligenza Artificiale e il livello di innovazione e sviluppo tecnologico dei Paesi d'origine. L'obiettivo è analizzare come variano i compensi dei ruoli più richiesti al variare del cluster di appartenenza, mettendo in relazione gli stipendi con la densità di annunci di lavoro e alcuni indicatori chiave, come ricerca, talento e infrastrutture.

Il confronto serve a capire se a un ecosistema tecnologico più maturo corrispondano stipendi più alti, o se la forte richiesta di competenze porti a retribuzioni elevate anche nei mercati emergenti. Si riescono, così, a evidenziare i legami tra le dinamiche retributive e il grado di sviluppo dei diversi cluster.

La dashboard in oggetto viene mostrata in Figura 5.2.

Grafici utilizzati

- *Bubble Matrix*: illustra la distribuzione dei primi quindici ruoli lavorativi per stipendio all'interno dei diversi cluster di sviluppo. L'ampiezza delle bolle è proporzionale

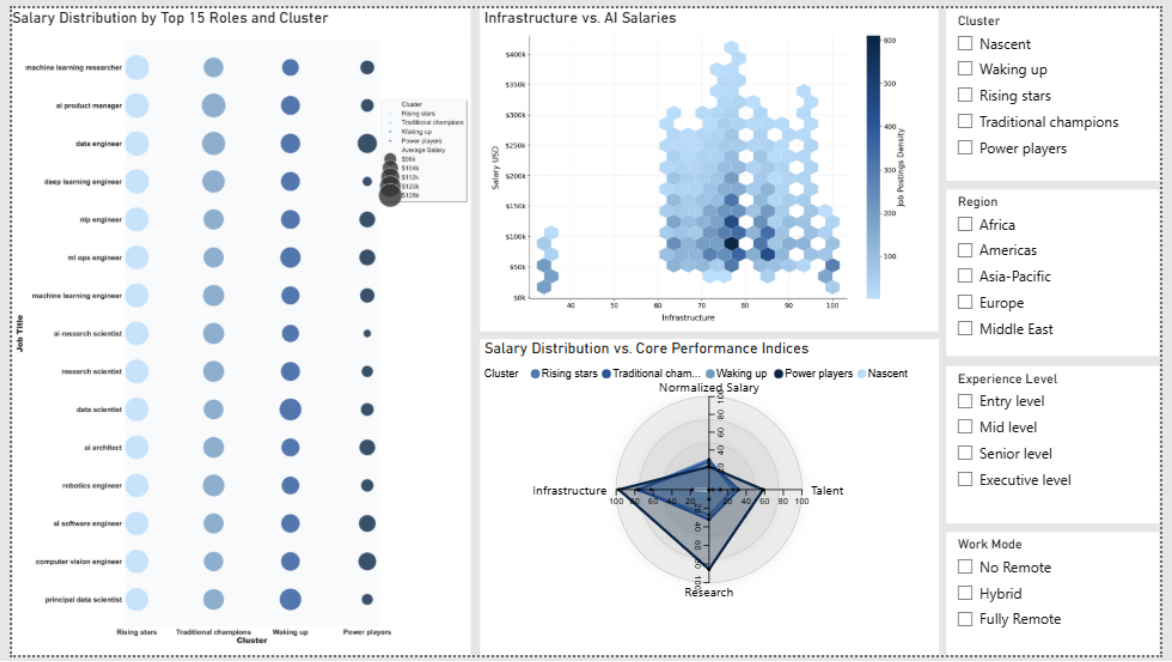


Figura 5.2: Global Salaries and Cluster Performance

allo stipendio medio, rendendo immediata l'individuazione delle figure retribuite maggiormente per ciascuna categoria. Questo grafico è stato realizzato tramite script Python, riportato nel Listato 5.1.

```
1
2       import matplotlib.pyplot as plt
3       import seaborn as sns
4       import matplotlib.ticker as ticker
5
6       plt.rcParams['figure.dpi'] = 300
7       plt.rcParams['savefig.dpi'] = 300
8       plt.figure(figsize=(18, 24))
9       sns.set_theme(style="white", rc={"axes.facecolor": "#
        f8fafc"})
10      my_blues = ["#badefb", "#749bc2", "#2a5599", "#0b2545"]
11      cmap = sns.color_palette(my_blues)
12
13      scatter = sns.scatterplot(
14          data=dataset,
15          x='Cluster',
16          y='Job Title',
17          size='Average Salary',
18          hue='Cluster',
19          sizes=(500, 5000),
20          alpha=0.8,
21          palette=cmap,
22          edgecolor='white',
23          linewidth=1.5
24      )
25
26      plt.xlabel('Cluster', fontsize=22, fontweight='800')
27      plt.ylabel('Job Title', fontsize=22, fontweight='800')
28
```

```
    plt.yticks(fontsize=20, fontweight='600')
    plt.xticks(fontsize=20,fontweight='600')

    formatter = ticker.FuncFormatter(lambda x, pos: f'${int(x/1000)}k')
    handles, labels = scatter.get_legend_handles_labels()

    new_labels = []
    for label in labels:
        try:
        val = float(label)
        new_labels.append(f'${int(val/1000)}k')
        except ValueError:
        new_labels.append(label)

    leg = plt.legend(
        handles,
        new_labels,
        bbox_to_anchor=(1.01, 0.9),
        loc='upper left',
        borderaxespad=0,
        fontsize=18,
        frameon=True,
        shadow=True
    )

    sns.despine(left=True, bottom=True)
    plt.subplots_adjust(left=0.45, right=0.99, top=0.99, bottom=0.1)
    plt.tight_layout()
    plt.show()
```

Listing 5.1: Script Python per la realizzazione della Bubble Matrix

- *Hexbin Plot*: mette in relazione il livello dell'infrastruttura tecnologica con gli stipendi erogati, dove la colorazione degli esagoni misura la concentrazione effettiva degli annunci di lavoro. Questo grafico è stato realizzato tramite script Python, riportato nel Listato 5.2.

```
    import matplotlib.pyplot as plt
    import pandas as pd
    from matplotlib.colors import LinearSegmentedColormap
    from matplotlib.ticker import FuncFormatter
    df = dataset
    my_colors = ["#BADEFB", "#749BC2", "#2A5599", "#0B2545"]

    my_cmap = LinearSegmentedColormap.from_list("custom_blue", my_colors)

    def salary_format(x, pos):
        return f'${int(x/1000)}k'

    fig, ax = plt.figure(figsize=(12, 8)), plt.gca()

```

```
15      hb = ax.hexbin(df['Infrastructure'], df['Salary USD'],
        gridsize=20, cmap=my_cmap, mincnt=1, edgecolors='white',
         linewidths=0.2)
16
17      cb = fig.colorbar(hb)
18      cb.set_label('Job Postings Density', fontsize=15)
19      cb.ax.tick_params(labelsize=13)
20      cb.outline.set_visible(False)
21
22      ax.set_xlabel('Infrastructure', fontsize=15, labelpad
        =10)
23      ax.set_ylabel('Salary USD', fontsize=15, labelpad=10)
24      ax.tick_params(axis='both', which='major', labelsize
        =13)
25      ax.yaxis.set_major_formatter(FuncFormatter(
        salary_format))
26
27      ax.grid(True, linestyle='--', alpha=0.3)
28      ax.spines['top'].set_visible(False)
29      ax.spines['right'].set_visible(False)
30
31      plt.tight_layout()
32      plt.show()
33
```

Listing 5.2: Script Python per la realizzazione dell'Hexbin Plot

- *Radar Chart*: permette di analizzare come il talento, l'infrastruttura e la ricerca influenzino lo stipendio medio per ogni cluster. Per consentire tale confronto, lo stipendio è stato normalizzato e portato su una scala da 0 a 100.

Conclusioni sull'analisi della Dashboard 6

L'analisi svolta nella sesta dashboard permette di osservare come lo sviluppo tecnologico dei singoli Paesi influenzi il mercato del lavoro dell'Intelligenza Artificiale.

La bubble matrix mostra che figure come Data Engineer, Machine Learning Researcher e Computer Vision Engineer mantengono compensi alti in tutti i cluster analizzati. Tuttavia, nei Paesi Power Players la media degli stipendi si abbassa leggermente rispetto ai cluster emergenti o intermedi, fenomeno dovuto probabilmente alla scarsità di profili qualificati che spinge le aziende a offrire ingaggi più alti per attrarre talenti.

Dall'hexbin plot si nota subito come il volume principale degli annunci di lavoro si concentri nei contesti con un livello di infrastruttura medio-alta (tra 70 e 85), all'interno di una fascia retributiva compresa tra $70.000 e $150.000. I valori degli stipendi più elevati, oltre i 300.000$, compaiono, invece, esclusivamente in presenza di infrastrutture con valore superiore a 75.

Infine, il radar chart mette in luce la struttura dei cluster: i Power player raggiungono valori elevati negli indici sull'infrastruttura e sulla ricerca (tra 80 e 100), mentre lo stipendio normalizzato rimane su livelli intermedi rispetto ai valori dei Traditional champion o dei Waking up. Questo può essere dovuto alla presenza di un buon numero di lavoratori con un basso livello di esperienza, il cui compenso fa diminuire la media retributiva complessiva.

5.2.3 Dashboard 7: Predictive Career and Market Growth

La settima dashboard analizza le relazioni tra anni di esperienza, livelli retributivi e indice globale dell'AI dei Paesi nel dataset. I KPI e i grafici integrati permettono di valutare il potenziale di crescita degli stipendi lungo la carriera e di verificare quanto lo sviluppo dell'ecosistema AI locale incida sui compensi reali.

La dashboard in oggetto è riportata in Figura 5.3.

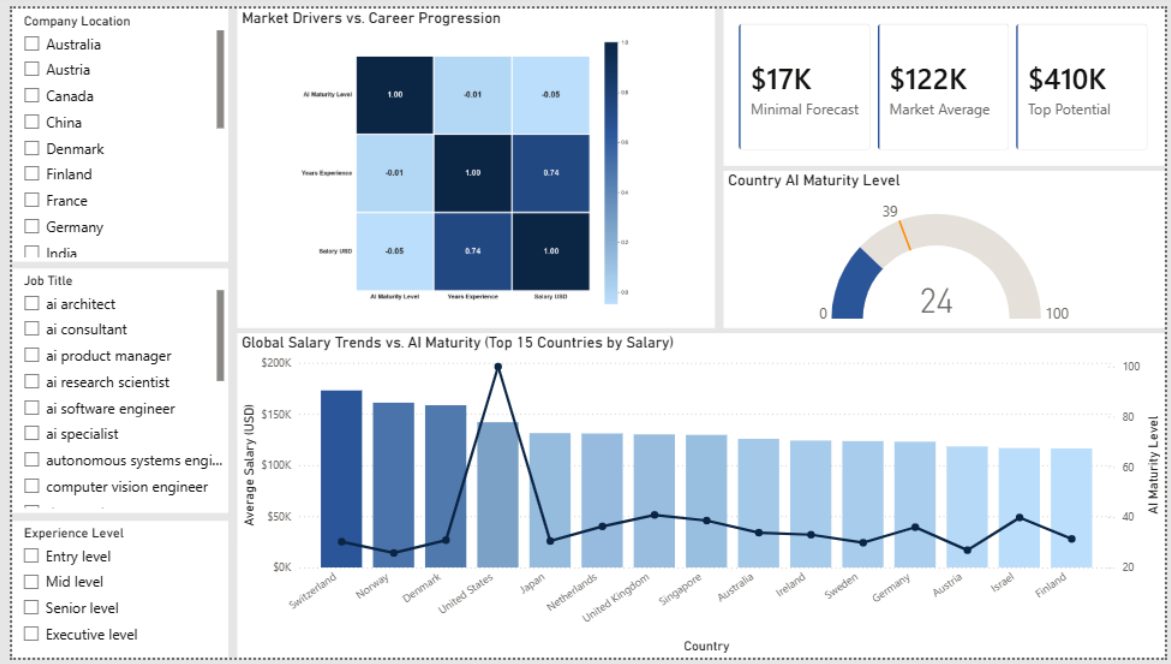


Figura 5.3: Predictive Career and Market Growth

Grafici utilizzati

- *Multi-row Card*: utilizzata per definire stime e obiettivi retributivi coerenti, evidenziando lo stipendio minimo (*Minimal Forecast* pari a $17.000), la media di mercato (*Market Average* pari a $122.000) e il valore massimo (*Top Potential* pari a $410.000).
- *Correlation Matrix*: calcola la correlazione tra anni di esperienza, stipendi e indice globale sull'AI. Rappresenta il riferimento principale per verificare in modo quantitativo l'impatto reale dell'esperienza sulla crescita retributiva. Questo grafico è stato realizzato tramite script Python, riportato nel Listato 5.3.

```
1   import matplotlib.pyplot as plt
2   import seaborn as sns
3   import pandas as pd
4   from matplotlib.colors import LinearSegmentedColormap
5   from matplotlib.ticker import FuncFormatter
6
7   df = dataset
8   my_colors = ["#BADEFB", "#749BC2", "#2A5599", "#0B2545"
    ]
9   df = df.rename(columns={'Total score': 'AI Maturity
    Level'})
10
11  corr = df.corr()
```

```
    my_cmap = LinearSegmentedColormap.from_list("
    custom_blue", my_colors)

    plt.figure(figsize=(10, 8))
    sns.set_style("white")

    ax = sns.heatmap(corr,
                     annot=True,
                     fmt=".2f",
                     cmap=my_cmap,
                     linewidths=1.5,
                     linecolor='white',
                     cbar=True,
                     square=True,
                     annot_kws={"size": 16, "weight": "bold"})

    plt.xticks(fontsize=12, weight='bold')
    plt.yticks(rotation=0, fontsize=12, weight='bold')
    plt.tight_layout(pad=2)
    plt.show()
```

Listing 5.3: Script Python per la realizzazione della Correlation Matrix

- *Misuratore*: mostra l'indice globale sull'AI, confrontando il valore del Paese selezionato con il target medio globale. Questo indicatore offre una panoramica macroeconomica utile a comprendere se il mercato di riferimento si trovi in una fase iniziale o sia già consolidato.
- *Grafico a linee e colonne*: mette a confronto la media degli stipendi dei vari Paesi con il rispettivo livello di maturità sull'AI. Esso serve a far emergere eventuali disallineamenti del mercato, evidenziando, ad esempio, le nazioni con retribuzioni elevate ma un indice tecnologico ancora contenuto, o, al contrario, i contesti in cui stipendi e sviluppo procedono in modo proporzionale.

Conclusioni sull'analisi della Dashboard 7

L'analisi svolta nella settima dashboard mette in evidenza, grazie alla multi-row card, un'ampia differenza tra lo stipendio minimo ($17.000), la media di mercato ($122.000) e il valore massimo ($410.000), confermando un'elevata eterogeneità e forti opportunità di crescita economica e professionale in questo mercato.

La matrice di correlazione relativa ai fattori di mercato e di carriera evidenzia come gli anni di esperienza rappresentino il fattore principale della crescita degli stipendi (0,74). Al contrario, il livello di maturità dell'AI del Paese mostra un impatto pressoché nullo sullo stipendio (-0,05), indicando che la retribuzione del professionista è guidata dalle competenze e dall'esperienza lavorativa piuttosto che dallo sviluppo dell'AI del Paese di riferimento.

Il grafico a linee e colonne mostra che i Paesi che offrono stipendi medi più elevati sono Svizzera, Norvegia e Danimarca. Tuttavia, si riscontra un'eccezione per gli Stati Uniti: pur posizionandosi leggermente al di sotto dei Paesi precedentemente elencati, gli USA hanno un indice di maturità nell'AI molto vicino al punteggio massimo.

Infine, il misuratore rivela un livello medio globale di maturità dell'AI di 24 su 100, ancora distante dal valore di riferimento pari a 39. Questo divario evidenzia come la maggior parte delle nazioni si trovi ancora nelle fasi iniziali o intermedie di adozione dell'AI, lasciando ampi margini di sviluppo futuro.

5.2.4 Dashboard 8: Market Dynamics and AI Readiness Landscape

L'ottava dashboard approfondisce la diffusione dell'AI a livello globale, confrontando i mercati attivi, caratterizzati dalla presenza di annunci di lavoro nel dataset, con quelli dormienti, privi di offerte registrate. Sfruttando la mappa, i KPI e i grafici di confronto sui vari pilastri, è possibile evidenziare dove si concentrano le differenze principali e valutare quanto le strategie dei singoli governi influenzino il mercato del lavoro.

La dashboard in oggetto viene mostrata in Figura 5.4.

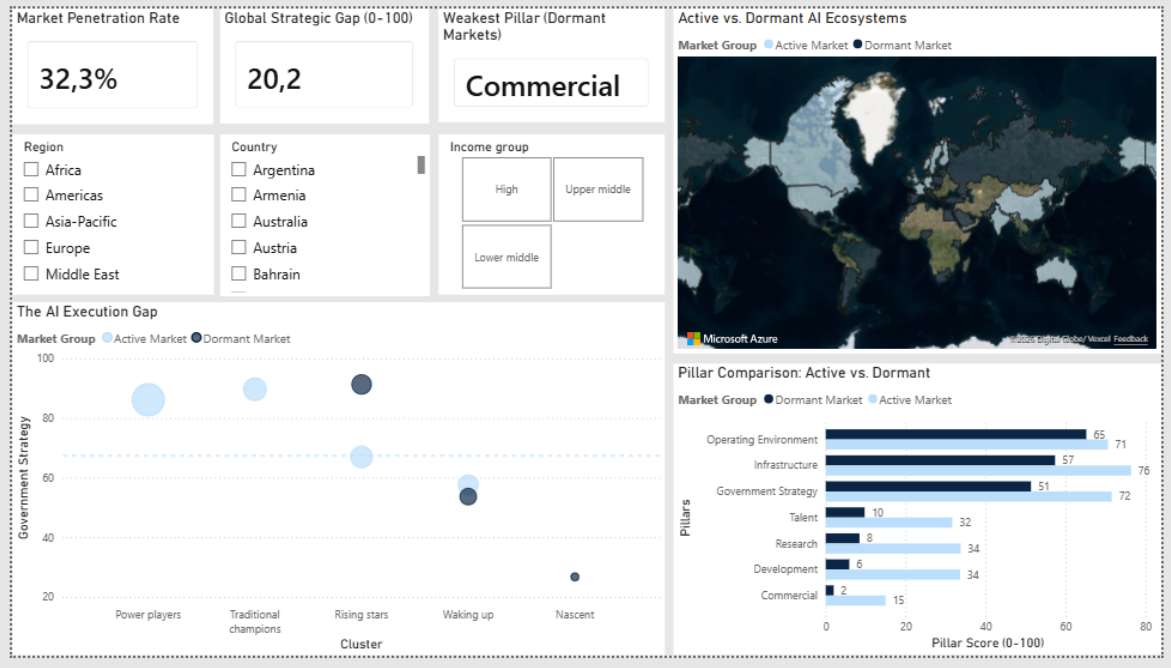


Figura 5.4: Market Dynamics and AI Readiness Landscape

Key Performance Indicator (KPI) utilizzati

- *Market Penetration Rate*: 32,3%; rappresenta la percentuale di Paesi con mercato attivo registrati nel dataset.
- *Global Strategic Gap (0-100)*: 20,2; indica la differenza dei valori delle strategie governative tra i Paesi con mercato attivo e quelli con mercato dormiente.
- *Weakest Pillar (Dormant Markets)*: Commercial; definisce l'indice con il punteggio medio più basso registrato tra i Paesi appartenenti ai mercati dormienti.

Grafici utilizzati

- *Mappa colorata*: permette di distinguere i Paesi attivi da quelli dormienti, mostrando come l'attivazione del mercato non sia distribuita in modo omogeneo ed evidenziando aree in cui il mercato rimane dormiente.
- *Grafico a dispersione*: mette in relazione i diversi cluster con l'indice Government Strategy, confrontando i mercati attivi e dormienti all'interno dei singoli cluster per evidenziare la differenza tra le strategie governative e la reale attività del mercato del lavoro.

- *Grafico a barre raggruppate*: confronta direttamente i due gruppi di Paesi lungo i diversi indici di analisi, evidenziando, per i mercati dormienti, le aree fondamentali che tali nazioni dovrebbero sviluppare.

Conclusioni sull'analisi della Dashboard 8

L'analisi dell'ottava dashboard offre una visione approfondita delle differenze tra i Paesi con mercato attivo e quelli con mercato dormiente.

Il valore del Global Strategic Gap, pari a 20,2, evidenzia una significativa differenza nell'applicazione di strategie governative tra questi due gruppi di Paesi, mostrando come l'impegno politico da solo non sia sufficiente a garantire un intervento della nazione nel mercato.

Il grafico a barre raggruppate, in linea con l'indice del Weakest Pillar, mette in luce il pilastro Commercial come la dimensione di maggiore debolezza per i mercati dormienti (con un punteggio medio di soli 2 punti), con ampi divari in Development, Research e Talent. Al contrario, i valori relativi a Operating Environment, Infrastructure e Government Strategy sono più simili tra i due gruppi.

Il grafico a dispersione mostra la presenza di Paesi dormienti con un alto Government Strategy nel cluster dei Rising star, dimostrando come le strategie governative non riescano a tradursi in offerte di lavoro concrete.

Infine, la mappa colorata illustra la distribuzione geografica di questo fenomeno, identificando l'Africa e il Sud America come aree prevalentemente caratterizzate da mercati dormienti.

CAPITOLO 6

Discussione

In questo capitolo viene presentata una discussione sull'intero percorso analitico sviluppato nel corso della tesi, descrivendo i principali punti di forza e le criticità riscontrate. Dopo aver affrontato le varie fasi del processo di Data Analytics, dalla fase di ETL alla realizzazione di dashboard interattive tramite Power BI, fino all'analisi descrittiva e diagnostica, risulta indispensabile valutare il contributo di tali attività alla comprensione del fenomeno analizzato, ma anche i limiti che hanno condizionato l'efficacia e la veridicità dei risultati.

6.1 Punti di forza dell'analisi condotta

6.1.1 Integrazione dei dati

Uno dei punti di forza principali dell'analisi condotta risiede nella capacità di integrare dati provenienti da origini diverse per ottenere un quadro con maggiori dettagli sull'impatto dell'Intelligenza Artificiale. Infatti, mettere in relazione dataset diversi ha permesso di far emergere correlazioni e insight che non sarebbero stati visibili analizzando ciascuno di essi in modo isolato.

6.1.2 Processo di ETL

La corretta impostazione del processo di ETL ha costituito il presupposto fondamentale per garantire l'affidabilità dell'intera analisi. Durante questa fase si è lavorato per correggere anomalie, gestire i valori mancanti e uniformare le tipologie di dato tra i diversi dataset. Questa fase di pulizia e trasformazione ha permesso di migliorare la qualità dei dati per prepararli alla realizzazione delle dashboard di analisi descrittiva e diagnostica.

6.1.3 Uso di Power BI

L'impiego di Power BI si è rivelato fondamentale per l'intero processo: dalla fase di ETL tramite l'editor Power Query, alla creazione delle relazioni tra i dataset in esame, fino alla definizione di misure e colonne calcolate in linguaggio DAX, per concludere con la realizzazione delle dashboard interattive.

6.1.4 Analisi di dashboard descrittive e diagnostiche

La realizzazione delle dashboard ha consentito di articolare l'analisi su due livelli complementari. La componente descrittiva ha fornito una panoramica immediata dei principali indicatori di sintesi, rendendo visibili andamenti e distribuzioni generali; a questa si è affiancato un approccio diagnostico, utile per approfondire le cause sottostanti a specifici risultati. L'uso combinato di filtri e relazioni tra i dati ha permesso, così, di isolare i principali fattori e interpretare con maggiore accuratezza le dinamiche del fenomeno.

6.2 Limiti e vulnerabilità dell'analisi condotta

6.2.1 Qualità dei dati

Uno dei limiti principali dell'analisi riguarda la qualità e la completezza dei dati di partenza. Anche se sono state usate fonti pubbliche autorevoli, ci si è scontrati con la dimensione ridotta dei campioni e con alcune incoerenze nelle informazioni raccolte. Attraverso la fase di ETL sono stati ripuliti i dataset e sistemati i valori mancanti, ma ciò ha rallentato l'intero processo. Tuttavia, tali interventi non hanno eliminato completamente la presenza di eventuali errori, rendendo necessaria una corretta interpretazione dei risultati ottenuti.

6.2.2 Difficoltà di allineamento dei dataset

Al fine di combinare le informazioni provenienti dai due dataset e condurre un'analisi più approfondita, si è cercato di metterli in relazione utilizzando l'identificatore primario del secondo dataset *Country* con la colonna *Company Location* del primo. Per verificare la presenza di corrispondenze ed evidenziare eventuali disallineamenti tra le fonti, è stata applicata un'operazione di *Left Anti Join*.

Da questa operazione sono emerse diverse criticità dovute alla differente provenienza dei dati. A causa di discrepanze nella nomenclatura e di alcuni valori non comuni, su un totale di 62 Paesi presenti in *AI Global Index*, 44 non hanno trovato alcun riscontro; tale scarto è stato successivamente ridotto tramite le operazioni di ETL implementate per uniformare i nomi delle nazioni.

L'elevato numero di elementi inizialmente privi di corrispondenza sottolinea la difficoltà di integrazione diretta tra le sorgenti, riducendo, di conseguenza, le possibilità di effettuare un confronto tra le variabili di entrambi i dataset.

6.2.3 Limiti temporali e copertura geografica dei dataset

Dall'analisi dei dataset sono emersi alcuni vincoli legati alla copertura geografica e al periodo temporale considerato.

Sul piano geografico, entrambi i dataset non includono la totalità dei Paesi mondiali. In particolare, il primo mappa un numero di nazioni inferiore rispetto al secondo, limitando la rilevazione dei dati ad alcune aree specifiche.

A questo si aggiunge una limitazione temporale propria del primo dataset, il quale raccoglie dati compresi tra l'inizio del 2024 e i primi mesi del 2025. Se da un lato questo lasso di tempo offre un quadro recente sulle dinamiche analizzate, dall'altro impedisce di svolgere analisi storiche di lungo periodo.

Conclusioni e sviluppi futuri

Il presente lavoro di tesi ha affrontato l'analisi dell'impatto dell'Intelligenza Artificiale sul mercato del lavoro globale, integrando una prospettiva incentrata su ruoli, livelli salariali, requisiti di istruzione e modalità lavorative (microscopica) e una focalizzata sugli indici di sviluppo, innovazione e investimenti nell'AI dei vari Paesi (macroscopica).

L'intero processo è stato condotto applicando una metodologia di Data Science, focalizzata principalmente sulla fase di analisi e supportata da Power BI come strumento di Business Intelligence. Attraverso l'elaborazione dei dataset *Global AI Job Market and Salary Trends 2025* e *AI Global Index* e l'implementazione di un processo di ETL (Extract, Transform, Load), è stato possibile normalizzare e integrare fonti di dati eterogenee in un modello relazionale solido. La modellazione dei dati, arricchita da misure e colonne calcolate in DAX, ha costituito la base per lo sviluppo di otto dashboard interattive da cui eseguire un'analisi descrittiva e un'analisi diagnostica.

L'intero percorso ha dimostrato l'importanza di disporre di dataset strutturati, aggiornati e affidabili, nonché di strumenti analitici in grado di trasformare i dati grezzi in informazioni utili al supporto decisionale.

L'impiego di Power BI ha assicurato la costruzione di un framework analitico estremamente flessibile. L'uso del linguaggio DAX per la logica di calcolo, unito all'integrazione di componenti di visualizzazione personalizzate, ha consentito di restituire dashboard ad alto impatto informativo, adatte a valutazioni di natura strategica.

Sotto il profilo professionale e formativo, l'attività svolta ha rappresentato un'importante occasione per consolidare e perfezionare le competenze acquisite nell'ambito della Data Science, della Business Intelligence e della gestione dei Big Data, con particolare riferimento all'implementazione del processo di ETL, alla modellazione relazionale dei dati e alla progettazione di dashboard interattive.

Tra gli sviluppi futuri più rilevanti potrebbe esserci l'adozione di strumenti di analisi predittiva e prescrittiva basati su modelli di Machine Learning, per stimare le evoluzioni salariali e l'emergere di nuove figure professionali nei prossimi anni.

Un'ulteriore linea di sviluppo potrebbe riguardare l'acquisizione di flussi di dati in tempo reale da piattaforme di assunzioni e fonti istituzionali, con l'obiettivo di rendere i modelli dinamici e adattabili ai rapidi mutamenti del settore tecnologico.

In conclusione, i risultati ottenuti confermano come la Data Science possa rappresentare un valido supporto strategico, offrendo strumenti efficaci per l'interpretazione dei dati riguardanti lo sviluppo e il mercato lavorativo dell'AI, ponendo così le basi per ulteriori approfondimenti sia in ambito accademico che aziendale.

Bibliografia

ALLOCCA, G. (2023), *Data analytics per tutti*, Apogeo, Milano.

BISWAS, N. e BISWAS, S. (2024), *Una guida completa alla ricerca sui processi ETL e alle tendenze future*, Edizioni Sapienza.

DI NUZZO, M. (2020), *Data Science e Machine Learning: Dai Dati alla Conoscenza*.

D'ONOFRIO, M. (2023), *DAX. Comprendere e usare il linguaggio delle formule di Microsoft Power BI*, Tecniche Nuove.

ERL, T., KHATTAK, W. e BUHLER, P. (2016), *Big Data Fundamentals: Concepts, Drivers & Techniques*, The Pearson Service Technology Series from Thomas Erl, Pearson.

FERRARI, A. e RUSSO, M. (2016), *Introducing Microsoft Power BI*, Microsoft Press.

KIMBALL, R. e CASERTA, J. (2004), *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*, John Wiley & Sons Inc.

OZDEMIR, S. (2017), *Data Science: guida ai principi e alle tecniche base della scienza dei dati*, Apogeo.

PASINI, P. e PEREGO, A. (2015), «Big Data: nuove fonti di conoscenza aziendale e nuovi modelli di management», Rapporto di ricerca, IBM.

SALVAGGIO, A. (2023), *Business Intelligence con Microsoft Power BI. Guida completa per l'analisi e la visualizzazione dei dati*, Edizioni LSWR, Milano.

YULIANTO, A. A. (2019), «APTIKOM Journal on Computer Science and Information Technologies», vol. 4 (2), p. 61–68.

ZAVARELLA, L. (2021), *Extending Power BI with Python and R: Perform advanced analysis using the power of analytical languages*, Packt Publishing.

Sitografia

- Akamai - https://www.akamai.com/it/glossary/what-is-big-data
- AlmaBetter - https://www.almabetter.com/bytes/articles/four-types-of-data-analysis
- Artera - https://www.artera.net/it/data-science/differenza-dati-strutturati/
- BMC - https://www.bmc.com/blogs/data-analytics-vs-data-analysis/
- Coupler.io - https://blog.coupler.io/power-bi-architecture/
- DataSkills - https://www.dataskills.it/le-tre-v-dei-big-data/#gref
- Databricks - https://www.databricks.com/it/blog/data-science-vs-data-analytics
- Deep Learning Italia - https://deeplearningitalia.com/corsi/corso-business-intelligence-strumenti-avanzati-power-bi-c0004/
- Everpure Data - https://www.everpuredata.com/it/knowledge/big-data.html
- Exigy - https://www.exigy.com/it/7-reasons-power-bi/
- Google Cloud - https://cloud.google.com/learn/what-is-big-data?hl=it
- IBM - https://www.ibm.com/it-it/think/topics/etl
- Jaro Education - https://www.jaroeducation.com/blog/demistifying-power-bi-architechture-and-features
- Kaggle - https://www.kaggle.com/datasets/bismasajjad/global-ai-job-market-and-salary-trends-2025?select=ai_job_dataset1.csv
- Kaggle - https://www.kaggle.com/datasets/katerynameleshenko/ai-index

- Microsoft Azure - `https://azure.microsoft.com/it-it/resources/cloud-computing-dictionary/what-is-data-science`
- Nextre - `https://www.nextre.it/cose-power-bi/#Per_gli_analisti`
- Opnet - `https://opnet.it/come-usare-i-big-data-in-azienda/`
- Opta - `https://www.opta.it/digital-transformation/business-intelligence/microsoft-power-bi/`
- Oracle - `https://www.oracle.com/it/big-data/what-is-big-data/`
- Osservatori Digital Innovation - `https://www.osservatori.net/blog/big-data-analytics/le-5v-dei-big-data/`
- SAS - `https://www.sas.com/it_it/insights/big-data/data-visualization.html`
- Talend - `https://www.talend.com/it/resources/what-is-data-analytics/`
- Visualitics - `https://visualitics.it/power-query/`
- Visualitics - `https://visualitics.it/dax-power-bi/`

Ringraziamenti

Desidero rivolgere un sincero ringraziamento alla mia famiglia per la presenza e il supporto garantito. Il loro sostegno ha rappresentato un importante punto di riferimento per il raggiungimento di questo traguardo.

Un ringraziamento va anche ai miei amici, per la vicinanza e per i momenti di spensieratezza vissuti insieme.

Infine, desidero ringraziare me stesso per la determinazione e l'impegno costante dimostrati nel superare le sfide di questo percorso e nel portare a termine questo lavoro.