POLITECNICO DI TORINO

Facoltà di Ingegneria

Corso di Laurea Magistrale in Ingegneria Gestionale

Tesi di Laurea Magistrale

L'impatto dei servizi cloud sulla diffusione dell'Artificial Intelligence: il caso Azure AI

Relatore

Prof. Fulvio Corno

Candidato

Umberto Ferrari

A.A. 2022/2023

Abstract

Il presente lavoro di tesi descrive l'analisi della diffusione dei servizi cloud di Artificial Intelligence, e nello specifico come l'offerta di tali prodotti stia modificando e potrà incentivare lo sfruttamento delle tecnologie di AI all'interno dei processi aziendali. La crescente attenzione da parte degli stakeholders a questa tematica ha portato le organizzazioni a soffermarsi sulle possibilità di integrazione di queste tecnologie all'interno dei propri processi, dalle quali consegue la necessità di comprendere le metodologie e le caratteristiche utili a supportare le imprese nella scelta di come sfruttarle nel breve e nel lungo periodo.
In questo contesto, l'elaborato mira a comprendere l'impatto dei servizi cloud nell'adozione dell'Artificial Intelligence e le possibilità di applicazione di tali processi per le organizzazioni, tramite l'esame dell'attuale mercato del cloud e l'approfondimento delle sfide e dei benefici legati all'utilizzo congiunto di queste due tecnologie, dettagliando le differenze tra le diverse architetture cloud.

L'obiettivo sarà quello di fornire una panoramica dei servizi AI offerti in cloud, comparando tra loro le tre metodologie di erogazione: IaaS (Infrastructure as a Service), PaaS (Platform as a Service) e SaaS (Software as a Service), in modo da comprendere i vantaggi e i rischi derivanti dall'adozione di queste tecnologie. Al fine di valutare questi parametri, la trattazione si concentrerà sui servizi offerti dalla piattaforma di cloud computing *Azure*, offerta da *Microsoft* e tramite la discussione delle fasi di progettazione e sviluppo di un applicativo basato sull'AI, realizzato secondo diverse infrastrutture cloud, sarà possibile analizzare i costi, diretti e indiretti dovuti all'uso di questa piattaforma, intesi come costi di servizio, sviluppo e manutenzione, e il time to market correlato alla realizzazione di tale prodotto, soffermandosi sull'impatto delle scelte architetturali su tali misure.

Indice

Elenco delle figure

Elenco delle tabelle

1 Introduzione

L'Intelligenza Artificiale (AI) è una delle tecnologie al momento più dibattute, offrendo opportunità per l'innovazione delle organizzazioni e la produttività degli individui.
In particolar modo, l'AI generativa ha avuto un impatto notevole, infatti, l'aumento della produttività per sviluppatori e lavoratori creativi che utilizzano sistemi di AI generativa, sta mostrando il suo potenziale e ha indotto le organizzazioni a riconsiderare i loro processi aziendali.

I modelli di AI hanno attraversato una rapida evoluzione nel corso degli ultimi anni: da ambito di ricerca si sono rapidamente trasformati in una tecnologia alla portata di tutti.
L'accelerazione dell'evoluzione dell'AI è stata guidata da fattori tecnologici e metodologici. Il notevole aumento della potenza di calcolo a nostra disposizione ha consentito l'addestramento e l'esecuzione di modelli di AI sempre più complessi in tempi considerevolmente ridotti. Allo stesso modo, l'espansione nella mole di dati disponibili è stato un fattore essenziale per l'addestramento di questi modelli. Infine, i progressi degli algoritmi di apprendimento automatico hanno consentito di estrarre e comprendere dati sempre più complessi e sofisticati.
Questi progressi hanno permesso di sviluppare e addestrare modelli di AI sempre più efficienti e precisi. Attualmente, in alcune specifiche applicazioni, l'AI supera persino le prestazioni degli esseri umani.

Grazie ai progressi ottenuti si è diffusa una maggiore consapevolezza delle potenzialità, così come dei rischi, nell'adottare questa tecnologia. Questo interesse non è più limitato ai soli processi di automazioni, ma anche e soprattutto, come strumento di veicolazione, assistenza e supporto alle attività manuali e creative. Tale diffusione si è manifestata in particolare nel contesto aziendale, dove lo sfruttamento dei servizi di AI può migliorare la produttività e affiancare nell'esecuzione dei task.

La diffusione capillare del servizio *ChatGPT* offerto da OpenAI, coadiuvata anche da un'ampia copertura mediatica, ha contribuito a dare l'impulso finale alla corsa all'AI che si è verificata nell'ultimo anno.
Il primato di *ChatGPT* è stato quello di essere il primo servizio di AI comprensibile dal grande pubblico, scalzando definitivamente l'AI dalla sua posizione di tecnologia elitaria dedicata a esperti del settore, un fenomeno di massa accessibile ai più.
Questo è stato un passaggio cruciale nella diffusione della tecnologia: sulla scia di ChatGPT sul mercato stanno nascendo sempre più servizi di AI dedicati alla risoluzione di specifiche attività, le aziende ne stanno riconoscendo le opportunità e le grandi società tech corrono per lanciare le proprie tecnologie proprietarie. Infatti, come riportato dal report *Standford AI Index 2023*[2]: *"Until 2014, most significant machine learning models were released by academia. Since then, industry has taken over. In 2022, there were 32 significant industry-produced machine learning models compared to just three produced by academia."* A dimostrazione di come le società tecnologiche stiano investendo sempre di più per lo sviluppo di questa tecnologia.

Tuttavia, l'adozione capillare dell'Intelligenza Artificiale è una sfida complessa per le aziende, rallentata da vari fattori:

- Avversione all'innovazione
- Scarsità di budget per creare e mantenere le infrastrutture IT necessarie
- Carenza di talenti tecnologici

Come rivelato da un sondaggio di Forbes[1], il 64% delle aziende ritiene che l'intelligenza artificiale contribuirà ad aumentare la produttività. Ciò conferma che, seppur con difficoltà, le aziende vedono un ampio potenziale nell'AI per migliorare i processi aziendali.

1.1 L'integrazione dell'AI e del cloud computing nel contesto tecnologico attuale

Uno sviluppo chiave che ha influenzato la diffusione delle tecnologie AI è stata l'adozione sempre più ampia dei servizi cloud e di come questi abbattano le barriere all'ingresso nel loro sfruttamento. Infatti, i provider di servizi cloud come *Amazon*, *Google* e *Microsoft* stanno ampliando la loro offerta di servizi in cloud per incentivare e semplificare la realizzazione di processi basati sull'Intelligenza Artificiale.
La possibilità di sfruttare risorse remote, scalabili e flessibili ha generato nuove opportunità per le aziende che intendono sfruttare l'intelligenza artificiale nei propri processi. L'uso delle tecnologie Cloud consente di implementare soluzioni basate sull'AI senza investimenti iniziali e complessità infrastrutturali significative.

Questi servizi stanno diventando noti come **AIaaS**: *Artificial Intelligence as a Service*, un termine coniato per classificare tutti i servizi che combinano l'AI con il modello di cloud computing.
Possiamo definire in generale l'integrazione tra AI e cloud computing, quindi, in una parola, AIaaS come: "Sistemi basati sul cloud che forniscono servizi on-demand per implementare, sviluppare, addestrare e utilizzare modelli di Intelligenza Artificiale"[3].
Da questa definizione possiamo riscontrare come il termine AIaaS non si riferisca solo ad applicazione in cloud di intelligenza artificiale ma che raggruppa servizi offerti con tutte le diverse tipologie di architetture cloud: IaaS (*Infrastrutture as a Service*), PaaS (*Platform as a Service*) e SaaS (*Software as a Service*). In linea di principio, i primi servizi disponibili sul mercato, e più diffusi, sfruttano il modello cloud convenzionale di tipo SaaS, ma con il diffondersi della tecnologia, i provider di servizi cloud hanno iniziato a proporre servizi di sviluppo di AI (PaaS) per permettere di personalizzare i modelli e servizi di infrastruttura di AI (IaaS) per permettere di realizzare e addestrare modelli customizzati in

cloud.
Possiamo suddividere integrazione dell'AI e del cloud in tre macro-blocchi che corrispondono alle tre tipologie di architettura esistenti nel modello cloud computing:

- Servizi software che offrono applicazioni che si basano su modelli di AI pronti all'uso, paragonabili al livello cloud SaaS
- Servizi di sviluppo di AI, identificabili come strumenti per supportare gli sviluppatori nel personalizzare i modelli di intelligenza artificiale e integrarli nelle proprie applicazioni e processi, paragonabili al livello cloud PaaS
- Servizi di infrastruttura di AI, che offrono un ambiente di sviluppo e una infrastruttura computazionale per la creazione e l'addestramento dei propri modelli di intelligenza artificiale, paragonabili al livello cloud IaaS

È interessante osservare come questi livelli non siano disgiunti tra loro, ma possono essere visualizzati come una pila di servizi correlati, posti uno sopra l'altro: il livello superiore si basa sui servizi di quello precedente, analogamente alle infrastrutture di cloud classiche.
Per esempio, un servizio SaaS si basa su uno o più servizi PaaS offerti dallo stesso provider, componendo un'infrastruttura interconnessa.

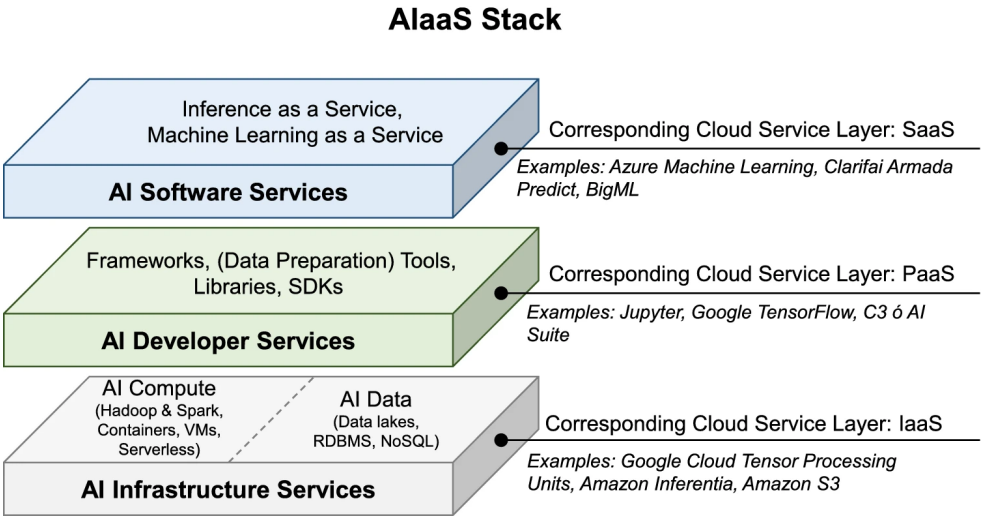

Figura 1: Rappresentazione dei servizi AI nelle infrastrutture cloud computing

L'obiettivo principale dell'integrazione tra AI e cloud computing è quello di rendere i modelli di intelligenza artificiali accessibili e convenienti, indipendentemente dalle dimensioni dell'impresa o dal suo budget. Il principio dello AIaaS è quello di fornire dei modelli pronti all'uso e permettere quindi agli utilizzatori di tralasciare lo sviluppo del modello e il suo allenamento e, invece, concentrarsi sull'integrazione dei servizi nel proprio processo di business.
I principali vantaggi nell'adozione dell'AI tramite il cloud computing sono i seguenti:

- Non necessita la realizzazione di un'infrastruttura IT e della gestione della sua manutenzione
- Evita di dover ricercare sul mercato o formare figure specifiche per sviluppare modelli di AI, riducendo le conoscenze specifiche richieste
- Permette di sfruttare una struttura scalabile, sia in termini di costi che di risorse disponibili

Analizzeremo in maniera più accurata nei prossimi capitoli i vantaggi e rischi nell'adozione dell'AI integrata con il cloud computing e l'impatto generato dalla scelta della tipologia di servizio sui costi, time to market e qualità del risultato.

1.2 Diffusione dell'AI nel panorama aziendale

L'adozione dell'AI da parte delle aziende è più che raddoppiata a partire dal 2017. Inoltre, anche il numero di funzionalità sfruttate, così come il budget investito nell'integrazione di processi di AI nella propria organizzazione ha mostrato un significativo aumento negli ultimi anni.

Dalla survey realizzata da *McKinsey & Company The state of AI in 2022 - and a half decade in review*[6] emerge come più della metà delle organizzazioni intervistate abbiano indicato di aver integrato l'AI in almeno una business unit o processo.

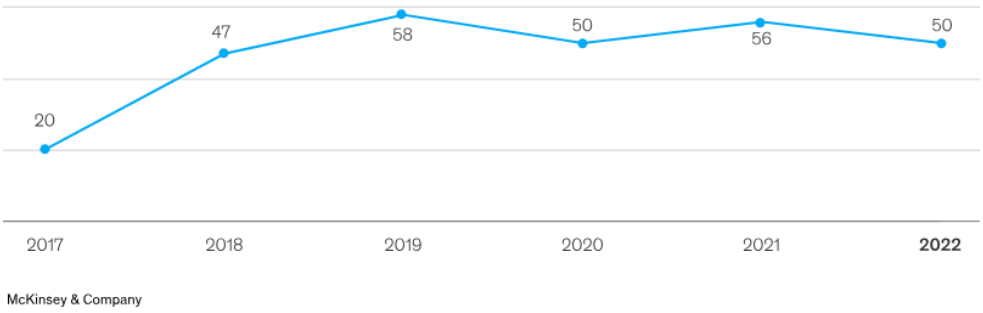

Figura 2: Andamento delle organizzazioni che hanno sfruttato l'AI, 2017 - 2022

Inoltre, è ancora più interessante notare come dalla stessa ricerca sia aumentato anche il numero medio di funzionalità di AI sfruttate da una stessa organizzazione, da una base di 1,9 funzionalità relative all'anno 2018, è emerso come questo valore sia raddoppiato a 3,8[6]. Questi due dati ci mostrano contemporaneamente un interesse sempre maggiore da parte delle aziende rispetto ai temi di AI, ma anche il fatto che quelle che avevano già sfruttato tali tecnologie abbiano incrementato il loro invisteminto per estenderla ad altri processi.

Ciò è confermato dal fatto che, grazie agli importanti sviluppi dell'AI, le aziende stanno sperimentando significative riduzioni dei costi e aumenti nei ritorni, nelle aree di business in cui hanno implementato l'AI per semplificare e automatizzare i processi.

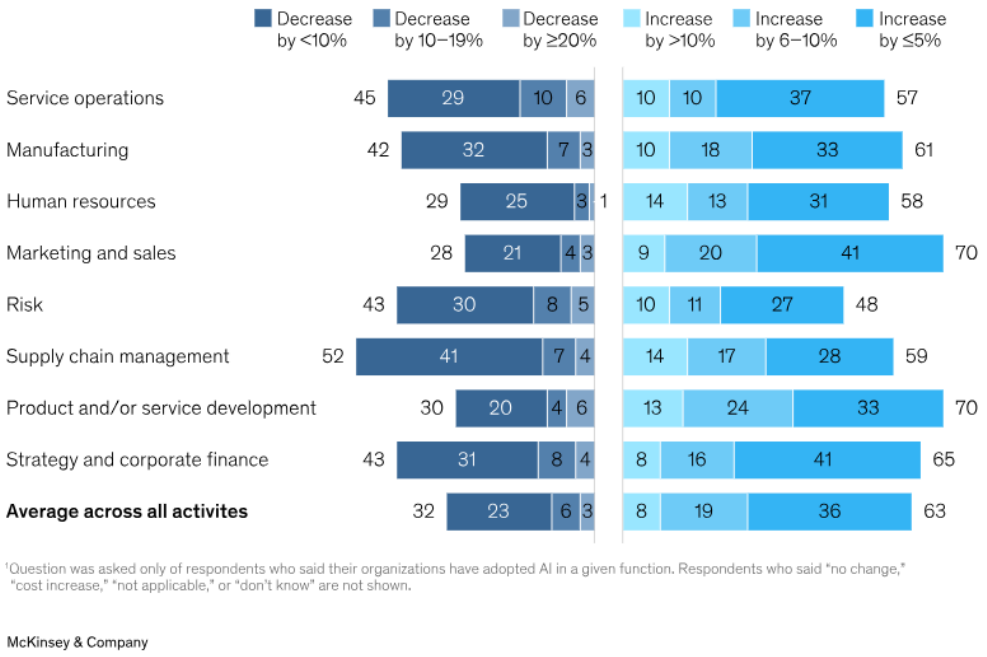

Figura 3: Riduzioni dei costi e aumento dei ritorni legati all'adozione dell'AI

I maggiori effetti sui ricavi si riscontrano nelle aree *Marketing and Sales* e *Product and/or service development*, aree ad alto tasso di creatività e di generazione di contenuti, in cui l'AI generativa può supportare e semplificare le attività. Invece, i maggiori risparmi si collocano nell'area della *Supply chain management*, in cui, già da diverso tempo, l'automazione della gestione da parte dell'intelligenza artificiale genera un forte impatto per l'impresa.

L'adozione dell'AI da parte delle aziende non è però immune da rischi: è stato riscontrato come molti dipendenti delle organizzazioni che hanno aperto all'uso dell'AI generativi inseriscano nei prompt informazioni personali o di proprietà dell'azienda. Da uno studio di *KPMG*[7] è emerso come un'alta percentuale di dipendenti che fanno uso di questi tool inseriscano informazioni private: riguardanti l'organizzazione per cui lavorano, informazioni finanziarie confidenziali o dati personali dei clienti.

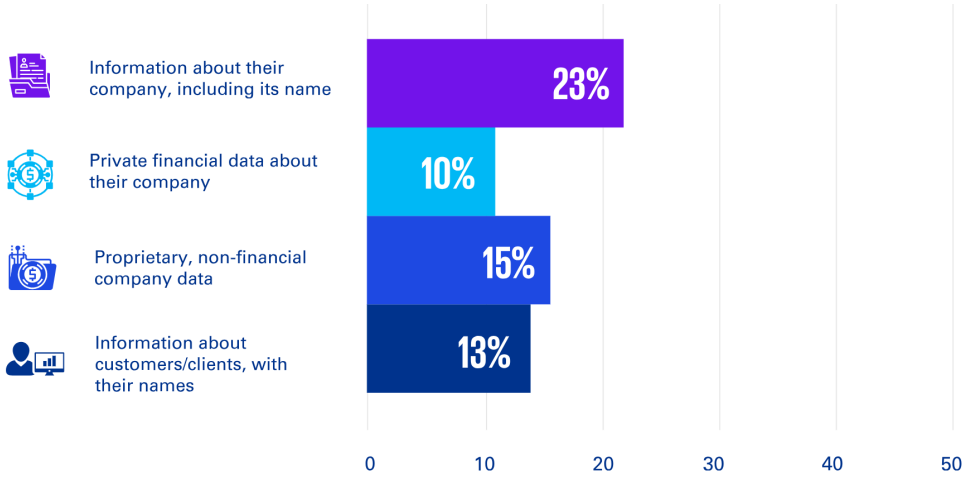

Figura 4: Informazioni confidenziali inserite nei prompt dei tool di AI generativa

Le informazioni inserite nei tool di AI pubblici sono spesso memorizzate dal provider per migliorare il modello o allenare le versioni successive dello stesso. C'è il rischio che il modello, addestrandosi, possa elaborare questa proprietà intellettuale per poi renderla disponibile a terzi, generando una fuga di dati.
La sicurezza delle informazioni e dei dati forniti ai modelli di AI sta diventando un tema centrale, svariate organizzazioni inizialmente aperte all'uso delle applicazioni di intelligenza artificiale da parte dei propri dipendenti, stanno vietando l'uso dei tool ad accesso pubblico a causa dei rischi inerenti alla sicurezza dei dati inseriti. La soluzione è quella di realizzare le proprie applicazioni di AI, in modo da controllare l'accesso ai tool, la sicurezza dei dati e la loro memorizzazione.

La diffusione dell'AI nelle aziende è trainata dal cloud, infatti come emerge dallo studio *IBM Global AI Adoption Index 2022*[4] solo l'8% delle aziende che sfrutta l'AI lo fa con un'infrastruttura on premises, quindi rinunciando al cloud. Per giunta, più della metà delle aziende che sfrutta il cloud lo fa tramite soluzioni public cloud o hybrid cloud.

Tabella 1: Ambienti cloud utilizzati dalle azienda per l'adozione dell'AI

Ambiente cloud	**% di utilizzo**
Private cloud	43%
Hybrid cloud o multicloud	32%
Public cloud	13%
On premises	8%

È infatti stimato che i ritorni nel mercato dei servizi di AI offerti con il modello del cloud computing quadruplicheranno nei prossimi 5 anni[5], passando dall'attuale valore di 51 miliardi di dollari a più di 200 miliardi di dollari, con pochi grandi player tech a contendersi il mercato.

2 Servizi cloud e intelligenza artificiale: una visione d'insieme

Nel seguente capitolo è riportata una panoramica dei servizi cloud, dei providers che offrono queste piattaforme, delle loro caratteristiche e i principali vantaggi e svantaggi. In particolare, saranno analizzate le diverse modalità di erogazione dei servizi da parte dei providers, già citate in precedenza.

Le conclusioni della seguente analisi costituiscono il contesto e i presupposti delle successive fasi di progettazione delle soluzioni proposte e delle scelte architetturali che saranno illustrate.

2.1 Definizione dei servizi cloud

Il cloud computing è un modello per consentire l'accesso, tramite un network, a un pool condiviso di risorse di elaborazione configurabili, fornite da un provider. Nella sua forma più diffusa, il cloud computing permette l'accesso tramite la rete internet a dei servizi, secondo tre tipologie di architettura, per sfruttarne le risorse messe a disposizione. Queste risorse condivise e distribuite utilizzano un'infrastruttura fornita da un provider di servizi cloud detto CSP.
L'architettura fisica del cloud computing prevede più datacenter dislocati sul globo con un modello per assicurare la ridondanza dei dati.

Esistono tre tipologie di cloud che si differenziano a seconda della modalità di accesso alle risorse e dal soggetto che le gestisce:

- **Public cloud**: è fornito da un CSP che possiede i data center e ne mette a disposizione le risorse sotto forma di servizi. Questi servizi sono messi a disposizione delle organizzazioni tramite la rete internet. In generale i cloud pubblici sono indipendenti dall'area geografica e altamente scalabili,

ma, a causa soprattutto dell'accesso tramite rete internet sono anche meno sicuri.

- **Private cloud**: è realizzato e configurato esclusivamente per un'organizzazione. Può essere fornito sia con architettura on premise, quindi con la realizzazione di un data center dedicato all'azienda, sia con architettura off premise, ovvero con la privatizzazione di una certa parte di risorse fornite da un CSP e rese inaccessibili al di fuori dell'organizzazione. I cloud privati hanno una scalabilità più limitata e tempi di rilascio delle risorse maggiori, ma, grazie all'accesso privato sono molto più sicuri.

- **Hybrid cloud**: rappresenta una combinazione tra public cloud e private cloud, in cui generalmente i servizi essenziali dell'organizzazione sono ospitare sulla parte privata e quelli non essenziali sulla parte pubblica. Questa tipologia risulta essere un buon compromesso tra sicurezza e scalabilità.

La scelta della tipologia di cloud è quindi legata a tre principali benchmark: costi, scalabilità desiderata, e livello di sicurezza richiesto. Ovviamente, le tipologie pubbliche risultano più economiche rispetto a quelle private, e di conseguenza più scalabili, ma il prezzo è un minor sicurezza delle informazioni, soprattutto nella fase di accesso e trasferimento tra utilizzatori del servizio e data center del provider.
Citiamo infine il termine **multicloud**, che non corrisponde ad una tipologia di cloud ma alla pratica diffusa nelle organizzazioni di sfruttare servizi cloud offerti da differenti CSP, nell'ottica di ridurre i costi o non legarsi esclusivamente a un unico fornitore.
In questo elaborato ci soffermeremo esclusivamente sulle tipologie public cloud e private cloud, considerando le restanti come casi particolari di quest'ultime.

Le principali caratteristiche del modello cloud computing sono:

- **Accesso onnipresente e on-demand alle risorse**: le risorse necessarie all'esecuzione delle operazioni richieste dall'utilizzatore vengono fornite in

automatico e su richiesta e l'accesso, tramite la rete (LAN o WAN) è assicurato a seconda delle esigenze dell'organizzazione.

- **Astrazione tra la risorsa sfruttata e l'infrastruttura sottostante**: l'utilizzatore dei servizi cloud non deve avere nessuna conoscenza o capacità per la gestione dell'infrastruttura su cui vengono eseguiti i servizi cloud; perciò, il cloud evita la necessità per le organizzazioni di avere a disposizione personale esperto su queste tecnologie.

- **Scalabilità delle risorse**, intesa come la capacità dei provider di distribuire rapidamente e su richiesta ulteriori risorse: i provider devono assicurare una rapidità nella distribuzione di ulteriori risorse agli utilizzatori, ad esempio per coprire i carichi di picco imprevisti.

- **Pagamento in base al consumo**: il pagamento dei servizi cloud è di solito gestito con un modello di *Pay as you go* ovvero l'azienda paga i servizi rispetto alla quantità di risorse che consuma.

In particolare, i principali vantaggi riscontrati dalle organizzazioni nell'adozione del cloud sono:

- **Riduzione dei costi fissi**: l'adozione del cloud permette di ridurre i costi fissi legati all'acquisto, alla gestione e alla manutenzione di un'infrastruttura IT on premise. Inoltre, l'utilizzo di impianti di questo tipo necessità di una serie di figure professionali specifiche, che tramite cloud possono essere evitate. Inoltre, la riduzione dei costi fissi permette un abbattimento delle barriere di ingresso alle tecnologie IT e quindi ne facilita lo sfruttamento.

- **Scalabilità**: le soluzioni basate sul cloud permettono alle organizzazioni di poter scalare le risorse in base alle esigenze e al carico di lavoro, avendo a disposizione una potenza di calcolo e uno spazio di storage idealmente illimitati. In un'infrastruttura on premise questo vantaggio sarebbe ottenibile esclusivamente con l'installazione di risorse in esubero, atte a gestire i pic-

chi del carico, generando ulteriori costi e non essendo comunque altrettanto rapido.

- **Rapidità**: i servizi cloud possono essere attivati in tempi brevissimi, ciò consente alle aziende di rilasciare nuove applicazioni o servizi più velocemente.

- **Sicurezza**: i CSP forniscono diverse tecniche di *data loss prevention*, come la ridondanza dei dati all'interno dello stesso data center e tra diversi data center distribuiti sul territorio. Inoltre, i data center prevedono sistemi di sicurezza avanzati per i rischi ambientali. In generale, la realizzazione di analoghe tecniche di sicurezza da parte delle organizzazioni stesse genererebbe considerevoli costi.

- **Innovazione**: le piattaforme cloud sono aggiornate in maniera continua e permettono di sfruttare le tecnologie più innovative in tempi rapidi.

Possiamo quindi riassumere i principali vantaggi dell'adozione del cloud con la semplificazione nella gestione delle risorse infrastrutturali e una conseguente facilità nello sfruttamento di ulteriori servizi, e una riduzione dei costi, soprattutto dei costi di ingresso. In generale, l'uso del cloud da parte delle aziende evita ingenti costi che invece sarebbero necessarie per la messa in opera di un'infrastruttura in sede, ciò permette, soprattutto a realtà di medie e piccole dimensioni di implementare soluzioni digitali.

Di contro però, l'adozione del cloud genera una serie di rischi e sfide per le organizzazioni:

- **Vendor lock-in**: o dipendenza dal fornitore, una volta adottata una piattaforma cloud offerta da uno specifico fornitore può diventare complesso e costoso migrare totalmente o in parte verso un diverso fornitore, a causa di vincoli tecnici tra i servizi offerti e i tempi necessari per riconfigurare le proprie applicazioni su una diversa piattaforma. Per questo motivo i

CSP detengono molto potere di mercato relativamente ai costi dei servizi, consapevoli che, una volta acquisito un cliente in maniera strutturale, sarà svantaggioso per quest'ultimo attuare un cambio di provider. Si tratta del principale rischio legato all'utilizzo del cloud, infatti molte organizzazioni per contrastarlo implementano un modello multicloud, accettando maggiori costi dovuti all'utilizzo contemporaneo di più piattaforme ma conservando una maggiore libertà nella scelta del fornitore per il futuro.

- **Limitato controllo**: l'infrastruttura su cui vengono eseguiti i servizi offerti così come i servizi stessi, sono di proprietà del fornitore che si occupa di gestirli e aggiornarli. Per questo motivo l'organizzazione non ne ha controllo diretto, ad esempio i CSP potrebbero decidere di ritirare dal mercato alcuni strumenti, obbligando alla migrazione verso altri che possano sopperire alla rimozione, o variarne unilateralmente le condizioni.

- **Accesso esclusivamente tramite internet**: il fatto che l'accesso alle risorse cloud sia esclusivamente tramite internet potrebbe generare dei problemi nel caso di disservizi momentanei della rete, in particolare non avendo accesso alle risorse fisiche in mancanza di connettività non è possibile eseguire nessuna operazione, nemmeno di monitoraggio.

- **Sicurezza**: oltre che un vantaggio è anche un rischio dell'adozione del cloud, infatti, seppur i provider attuino tra le migliori tecnologie di sicurezza per impedire l'indisponibilità dei servizi o la perdita di dati, sono comunque presenti dei rischi dovuti alla possibilità che i dati vengano violati.

Per concludere, il cloud computing, come tutto le tecnologie, comporta una serie di vantaggi ma anche di rischi, per questo motivo le organizzazioni devono valutare i propri benefici nell'adozione del cloud, in particolare se la riduzione dei costi ottenuta sopperisce ai rischi che è necessario accettare.

2.2 Architetture cloud: IaaS, PaaS, SaaS

Il cloud computing può essere suddiviso in tre principali modelli a seconda della modalità erogazione dei servizi:

- **Software as a Service (SaaS)**: forniscono un'applicazione configurabile sotto forma di servizio
- **Platform as a Service (PaaS)**: forniscono un ambiente di personalizzazione e sviluppo, accessibile tramite un'interfaccia web e sfruttabile tramite delle API integrabili nelle applicazioni
- **Infrastructure as a Service (IaaS)**: forniscono un'infrastruttura delocalizzata, compresa di spazio di memorizzazione e potenza di calcolo per ospitare i propri servizi o applicazioni

Nei seguenti paragrafi analizzeremo nello specifico ogni modello di erogazione, mettendone in luce i principali vantaggi e rischi.

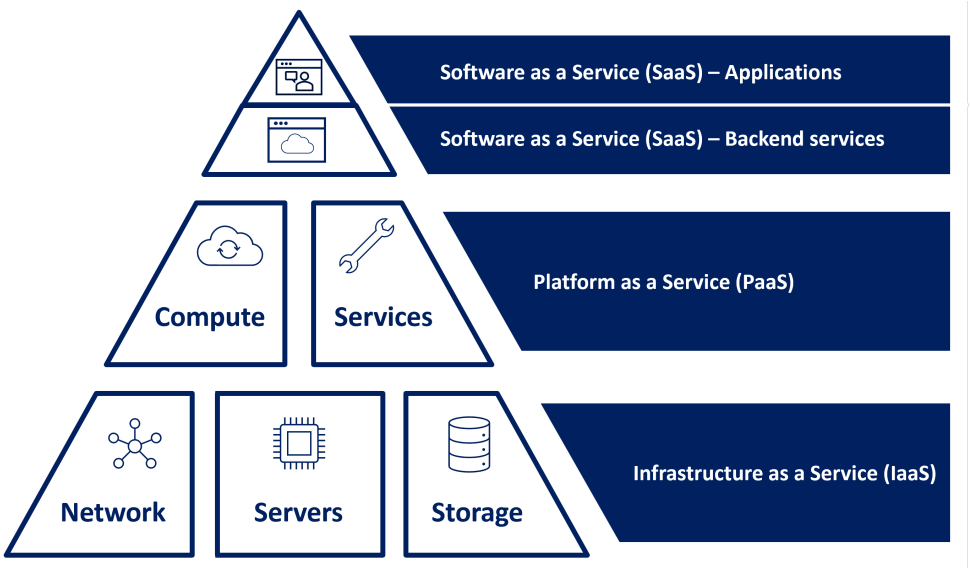

Figura 5: Stack delle tipologie di erogazione dei servizi cloud

2.2.1 Software as a Service

Software as a Service (SaaS) è il modello di erogazione posto alla cima dello stack dell'architettura cloud[5], prevede l'offerta di un servizio sotto forma di applicazione, questa può prevedere una web application o strumenti utilizzabili lato backend nei propri processi. Il provider fornisce l'infrastruttura e le piattaforme necessarie all'utilizzo dell'applicazione, così come i metodi di sicurezza e di integrazione in altre applicazioni o servizi, e si impegna a mantenere, aggiornare e assicurare l'accesso alla piattaforma.
I servizi SaaS appaiono al cliente come un applicativo web-based presentato tramite un'interfaccia accessibile mediante la rete internet.
SaaS è il modello più utilizzato e che genere i maggiori introiti per i fornitori. Ciò è dovuto principalmente alla sua massima accessibilità. Infatti, molte applicazioni, ampiamenti diffuse, sono offerte con il paradigma SaaS, per esempio i servizi di e-mail come Gmail e Outlook, Microsoft 365 nella parte online o Microsoft Teams.

L'utente finale dei servizi di questo tipo può utilizzarli direttamente *out of the box* provvedendo solo a caricare nel servizio i propri dati, oppure, alcuni strumenti SaaS permettono un certo grado di personalizzazione, secondo delle opzioni di configurazione predefinite dal provider. A differenza dei tradizionali software, l'utente non deve gestire:

- L'acquisto di una licenza, sostituita dal pagamento di un abbonamento flessibile
- L'installazione del software e l'acquisto dell'hardware su cui eseguirlo
- L'aggiornamento e la manutenzione del software e dell'hardware annesso

Inoltre, con il paradigma SaaS il cliente si assicura un accesso onnipresente e multiutente all'applicativo.

I principali rischi del modello SaaS si collocano principalmente a livello della protezione dei dati e alla dipendenza dal fornitore: i dati, essendo trasferiti tramite rete internet, e memorizzati nell'infrastruttura del provider non sono fisicamente gestiti dal cliente, perciò c'è il rischio che vadano persi, violati o che risultino inaccessibili a causa di criticità nella rete o nei data center del fornitore; inoltre, a seguito della diffusione di un'applicazione SaaS in un'organizzazione la migrazione verso strumenti concorrenti o proprietari può risultare complessa e costosa, perciò spesso i clienti rimangono legati ad uno stesso fornitore nel tempo.

Il costo degli strumenti forniti a modello SaaS è stabilito secondo tre alternative:

- Accesso completo e illimitato all'applicativo a fronte di un abbonamento
- Accesso all'applicazione a fronte del pagamento di un abbonamento rapportato al numero di utenti
- Accesso all'applicazione a fronte del pagamento di un abbonamento rapportato alle risorse consumate

I costi finali possono essere molto variabili, una stessa applicazione può essere fornita anche con diverse alternative nel calcolo del valore dell'abbonamento, ma soprattutto il costo totale è dettato dalla tipologia di applicazione che si intende sfruttare.

La scelta di uno strumento erogato secondo il metodo SaaS risulta quella più indicata per le organizzazioni che intendono sfruttare un applicativo senza attuare una considerevole attività di sviluppo e configurazione, potendolo utilizzare direttamente come viene offerto dal provider.

2.2.2 Platform as a Service

Platform as a Service (PaaS) fornisce un ambiente di sviluppo tramite il quale è possibile realizzare e rilasciare la propria applicazione per distribuirla tramite via internet senza la necessità di dover gestire l'infrastruttura sottostante. Costituisce il livello intermedio dello stack dell'architettura cloud[5], rappresenta il punto di accesso alle risorse offerte e la base delle applicazioni a paradigma SaaS. Il cliente non deve occuparsi della gestione fisica o virtuale dell'infrastruttura, il provider la fornisce in modalità on-demand per lo sviluppo e l'esecuzione dei software distribuiti, intesa come potenza di calcolo, rete e storage. Tramite il paradigma PaaS i provider offrono anche le proprie soluzioni software altamente personalizzabili, senza limitazioni predeterminate. Infatti, mediamente i servizi PaaS le organizzazioni possono realizzare le proprie soluzioni SaaS e distribuirle, sia internamente sia verso terzi.

I servizi PaaS generano per l'utente vari vantaggi:

- Scalabilità dell'applicazione rilasciata
- Distribuzione semplificata di nuove versioni
- Componenti necessarie alla distribuzione di un'applicazione come: sicurezza, autenticazione, gestione del carico e data retention sono comprese nel servizio acquistato e quindi non richiedono uno sviluppo dedicato
- Possibilità di sfruttare software e processi sviluppati da terzi ma completamente personalizzabili

Dall'altro lato, i rischi connessi allo sfruttamento di questo modello non si discostano dal precedente, ovvero la sicurezza dei dati e in questo caso anche della disponibilità dell'applicazione rilasciata: non avendo accesso all'infrastruttura eventuali disservizi non possono essere mitigati dal cliente. Invece, per quanto riguarda il rischio di dipendenza dal fornitore, pur essendo considerevole, grazie

al maggior grado di personalizzazione dello strumento, la migrazione risulta meno impattante.

I costi dei servizi PaaS, oltre che alla tipologia di servizio, sono commisurati al consumo che ne viene fatto, in altre parole, calcolati in base ai tempi di esecuzione e alla memoria occupata. In altri casi, il provider prevede dei tiers di servizio, che corrispondono a un certo quantitativo massimo di consumo di risorse in un arco di tempo prestabilito, a cui corrisponderà un prezzo fissato a priori. Per questo motivi, i servizi PaaS risultano i più flessibili in termini di costi, soprattutto nel caso di un utilizzo limitato o saltuario.
Infatti, i servizi PaaS sono indicati per le realtà che intendono realizzare i propri processi, effettuando anche attività di sviluppo considerevole, senza dover gestire l'infrastruttura necessaria, e soprattutto nei casi in cui si prevede un carico più ridotto o discontinuo nel tempo delle risorse.

2.2.3 Infrastructure as a Service

Infrastructure as a Service (IaaS) costituisce le fondamenta dell'architettura cloud[5], prevede la fornitura di risorse di archiviazione e di calcolo utilizzate da sviluppatori per distribuire soluzioni proprietarie. Sulla base di servizi IaaS, il cliente può realizzare ed eseguire i propri servizi, che possono costituirsi anche di strumenti a paradigma PaaS o SaaS.
I provider offrono le risorse richieste e le interfacce amministrative utili a gestire e monitorare queste quest'ultime. Mediante i servizi IaaS, un'organizzazione può configurare in breve tempo un'infrastruttura complessa, che ha il proprio nodo centrale in una o più virtual machine (VM), ovvero dei server virtuali annessi ad una data potenza di calcolo e spazio di memorizzazione.

Il provider è responsabile della gestione, manutenzione e della sicurezza dell'hardware, però il cliente è in grado di controllarlo ampiamente: settare l'indirizzo IP, il sistema operativo, i servizi distribuiti su di esso, le prestazioni e può connet-

terlo alla propria rete privata tramite una VPN per farlo apparire come parte della propria infrastruttura. Le risorse hardware richieste dal cliente vengono dedicate esclusivamente a quest'ultimo e non sono condivise per altri utilizzi, a differenza delle risorse consumate per i servizi offerti secondo le altre metodologie.

L'utilizzo di modelli IaaS offre il vantaggio di poter utilizzare un'infrastruttura completa e scalabile senza doversi sobbarcare i costi correlati all'installazione e gestione dell'hardware. Di contro però, l'impossibilità di poter accedere fisicamente alle risorse potrebbe impedire al cliente di effettuare alcune operazioni, soprattutto in caso di disservizio o di perdite di dati.

I costi dei servizi offerti a modello IaaS sono correlati alla quantità di risorse richieste, in termini di CPU, RAM e spazio di memorizzazione, oltre che alle richieste al contorno come sistema operativo o particolari configurazioni di rete. In generale, i prezzi degli strumenti IaaS risultano meno flessibili rispetto alle altre tipologie di erogazione, prevedendo sostanzialmente una scelta delle risorse che si ripercuote su un costo fissato. Per questo motivo risultano in molti casi meno vantaggiosi in termini economici rispetto alle altre proposte, non prevedendo dei pagamenti in base al consumo o in base al tempo di utilizzo.
Le organizzazioni scelgono di adottare dei servizi IaaS quando intendono sviluppare degli applicativi proprietari e quindi improntate ad elevate attività di sviluppo e per le quali si prevedono carichi elevati e continui per cui risulta vantaggioso l'acquisto di risorse dedicate.

2.3 Cloud service providers

In questo paragrafo si analizzerà il mercato dei servizi cloud, evidenziato i principali provider presenti e mettendo in luca la posizione che la piattaforma *Microsoft Azure* ricopre in esso.

Il mercato dei servizi cloud si caratterizza da offerte standardizzate e altamente automatizzate, fornite da gestori detti cloud service providers (CSP). Al fine di fornire una visione d'insieme dei principali fornitori possiamo prendere a riferimento lo studio realizzato da Gartner[13]: *Magic Quadrant for Cloud Infrastructure and Platform Services*[8]. Il report, realizzato a fine 2022, riporta una panoramica dei player presenti sul mercato e del loro posizionamento secondo due parametri:

- **Ability to Execute**: valuta la realizzazione del servizio, prendendo in considerazione: la qualità del prodotto offerto, la redditività complessiva generata, la reattività sul mercato intesa come capacità di adattarsi al mercato e il ruolo delle vendite e del marketing e l'esperienza testimoniata dai clienti.
- **Completeness of Vision**: insieme di criteri che valutano la completezza dell'offerta e della strategia, tra cui: comprensione del mercato, strategie di marketing e delle vendite, il grado di innovazione e la strategia nella fornitura dei servizi.

Il risultato dello studio è riassunto graficamente nel *Magic Quadrant*[6], suddiviso in quattro aree in cui si collocano gli attori del mercato, identificati sulla base dei valori assunti dai parametri sotto riportati:

- **Leader**: si caratterizzato dall'offerta di servizi adatti ad una vasta gamma di applicazioni e rapidamente adattabili alle esigenze
- **Challengers**: si caratterizzato dall'offerta di una gamma limitata servizi dedicati a soddisfare specifiche esigenze

- **Visionaries**: si tratta di fornitori emergenti ma che stanno realizzando significativi investimenti per ampliare la loro offerta
- **Niche players**: si tratta degli attori di nicchia del mercato, dedicati a specifiche applicazioni o regioni in cui operano

Nel *Magic Quadrant* rientrano otto fornitori di servizi cloud, di cui tre si collocano nell'area dei Leaders: *Amazon Web Services*, *Google Cloud Platform* e *Microsoft Azure*.

Figura 6: Magic Quadrant dei cloud service providers

2.3.1 Amazon Web Services

Amazon Web Services (AWS)[14] è un ampio fornitore di servizi cloud, concentrato sull'espansione del mercato verso nuovi ambiti. AWS propone la più profonda gamma di soluzioni offerte, stabilendo anche un ruolo guida nel settore, sviluppando metodologie che si trasformano spesso in standard che si diffondono anche ai competitors. Infatti, nella valutazione di *Gartner* ha ottenuto il punteggio maggiore in termini di *Ability to execute*.
AWS si colloca come incontrastato leader del mercato in termini di entrate, superando di due volte *Microsoft Azure*, suo principale concorrente. Grazie alla sua ampia diffusione rappresenta la scelta primaria per altri attori del settore che vogliono offrire le proprie soluzioni cloud-based.
Di contro, però, a causa della sua posizione dominante AWS propone una debole strategia a supporti dei clienti che intendono adottare soluzioni multicloud o proprietarie. Inoltre, ha una presenza territoriale, intesa come collocazione dei datacenters, limitata sul continente l'Europa.

In tema AI, AWS offre prodotti di AI e ML con modello IaaS e PaaS, con la possibilità di utilizzare modelli proprietari, sfruttare modelli personalizzati da terzi e offerti sul marketplace oppure customizzare il proprio modello.

2.3.2 Google Cloud Platform

Google Cloud Platform (GCP)[15] ha visto una significativa crescita negli ultimi anni, investendo soprattutto nella propria offerta di servizi IaaS e PaaS. I clienti che scelgono GCP tendono ad essere start-up o imprese nativamente cloud-based. GDP ha registrato il più alto incremento nei ricavi rispetto a qualsiasi altro player analizzato, soprattutto grazie a un miglioramento della comunicazione verso i clienti, in modo da mutare la reputazione di Google come fornitore IT inadatto al contesto aziendale. Al contrario, GDP risulta ancora l'unico fornitore preso in esame a registrare delle perdite finanziarie.

Sul tema Artificial Intelligence la piattaforma di Google è maggiormente incentrata sull'offerta di servizi per l'analytics dei dati raccolti, e più in generale, come AWS offre soluzioni di tipo IaaS e PaaS con strumenti proprietari.

2.3.3 Microsoft Azure

Microsoft Azure[16] si colloca come secondo principale fornitore del mercato, e risulta la soluzione particolarmente adatta alle organizzazioni incentrate sui prodotti Microsoft, per le quali permette un'integrazione semplificata con la preesistente infrastruttura. Infatti, i suoi clienti tendono ad essere aziende di medie e grandi dimensioni, che attuano anche una migrazione verso il cloud.
Analizzando il *Magic Quadrant* emerge come Microsoft abbia ottenuto la migliore valutazione in termini di *Completeness of vision* a testimonianza degli importanti investimenti nel miglioramento dell'architettura e dei servizi offerti; infatti, Azure è orientato alle soluzioni offerte e grazie a ampie collaborazioni con altri attori del settore permette di soddisfare la totalità dei casi d'uso dei clienti. Inoltre, Microsoft offre il miglior approccio nella strategia multicloud e hybrid cloud, propone servizi di gestione e governance per semplificare le operazioni di gestione di infrastrutture cloud di questo tipo.

Microsoft si colloca nella posizione di fornitore più diffuso per i servizi a modello SaaS, infatti con la piattaforma Power Platform, parte del proprio ecosistema cloud, offre una vasta gamma di servizi per la realizzazione di web applications e processi di automazione, anche basati sull'utilizzo dell'artificial intelligence, adatti a molteplici utilizzi.

Azure risulta anche una piattaforma in forte crescita, seconda solo a AWS, ma con un divario molto più limitato in Europa. Microsoft sfrutta la sua posizione dominante relativa all'utilizzo delle licenze dei suoi prodotti software come Windows e SQL Server, in modo da disincentivare l'utilizzo di questi prodotti con fornitori concorrenti. Inoltre, si è riscontrata una complessità significativa per i

clienti di Azure nella gestione e previsione dei costi, lamentando costi imprevisti e difficoltà a consolidare la propria spesa.

Infine, è necessario riportare la partnership intrapresa da Microsoft con OpenAI[17] a inizio 2023[9], per ovvi motivi temporali questa novità non è stata presa in esame nello studio analizzato, ma conferma gli consistenti investimenti attuati da Microsoft per proporre soluzioni realizzate da altri player del settore IT. Grazie a questa collaborazione è possibile sfruttare servizi cloud integrati nell'ecosistema Azure per realizzare soluzioni basate sui modelli di AI sviluppati da OpenAI come DALL-E e GPT.
Attualmente, Microsoft risulta l'unico provider ad offrire servizi in partnership con OpenAI, oltre che le proprie soluzioni di AI. Inoltre, tra i fornitori citati è l'unico incentrato anche su soluzioni SaaS pronte all'uso, anch'esse che sfruttano non solo tecnologie proprietarie ma anche modelli realizzati da partner di Microsoft.

2.3.4 Conclusioni sul mercato del cloud

I tre player analizzati, le *Big Three* nel mercato del cloud, si contendono la maggior parte della quota complessiva globale.
Il cloud è un mercato in forte crescita, secondo le ultime analisi di *Gartner*, si stima che nel 2023 si registrerà un incremento del 21,7% nella spesa da parte delle organizzazioni per i servizi cloud, generando una spesa globale di 597,3 miliardi di $, rispetto ai 491 miliardi del 2022[10]. La maggiore crescita sarà riscontrata nei segmenti IaaS e PaaS, trainati dalla diffusione dell'intelligenza artificiale, in particolare dei *Large Language Model (LLM)*. Infatti, come emerge dalla medesima ricerca, si prevede che entro il 2026 il 75% delle organizzazioni adotterà il cloud come piattaforma IT di riferimento.
Analizzando la ripartizione del mercato tra i principali providers emerge che i 3 fornitori analizzati coprano il 68% del mercato globale dei servizi cloud[11].

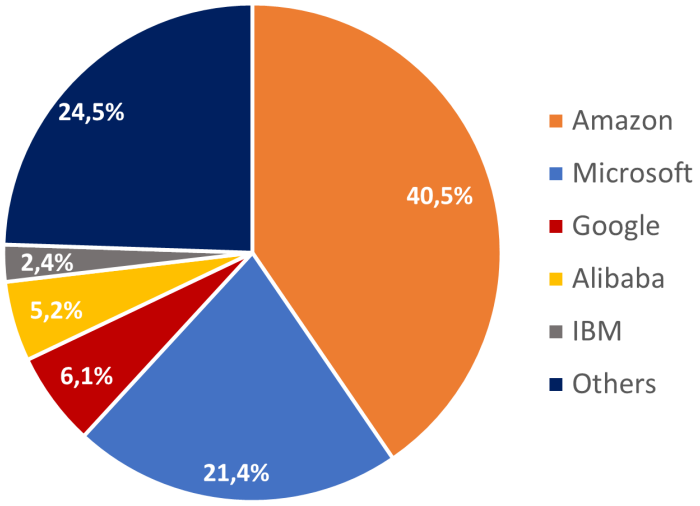

Figura 7: Market shares dei cloud service providers (esclusi servizi *SaaS - Applications*)

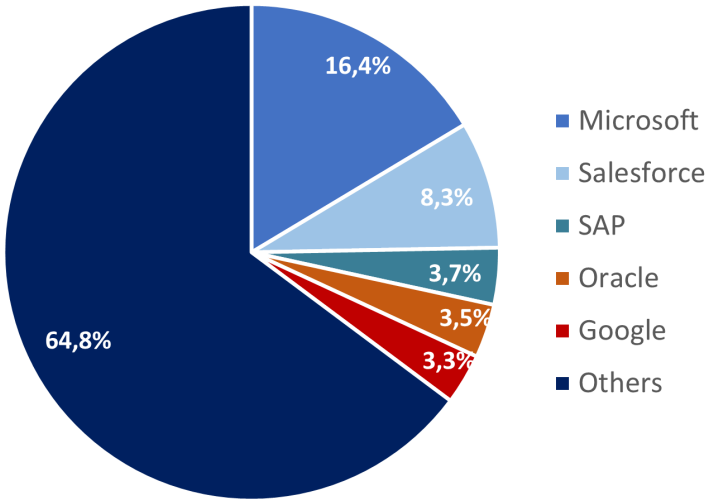

Figura 8: Market shares dei cloud service providers (esclusivamente servizi *SaaS - Applications*)

Microsoft, con la piattaforma Azure, detiene una quota del 21,4% posizionandosi come secondo player del mercato. Se si prendono in considerazione esclusivamente i servizi di tipo *SaaS - Applications*, risulta invece il principale player del mercato con una quota del 16,4%, in questa seconda analisi la piattaforma AWS

non è considerata tra i maggiori provider. La posizione di Azure risulta ancora più favorevole analizzando il solo mercato Europeo.
Al momento della stesura del presente elaborato non erano disponibili studi che mettessero in luce la percentuale di adozione di soluzioni AI cloud-based per i principali provider.

Concentrando l'attenzione sulle funzionalità offerte, a discapito della minore quota di mercato rispetto ad AWS, Microsoft si identifica come un fondamentale attore del settore. L'offerta della piattaforma è completa per tutte le applicazioni e casi d'uso, integrando un importante supporto alle strategie multicloud. In particolare, risulta essere il fornitore con la gamma di servizi che coprono tutti i modelli di erogazioni, concentrandosi ampiamente anche sulle soluzioni SaaS.
Infine, grazie alla partnership con OpenAI, attualmente Azure rappresenta il migliore punto di accesso alle tecnologie di AI più innovative e richieste, le quali, insieme alle tecnologie proprietarie permettono di semplificare e stimolare l'integrazione dell'intelligenza artificiale da parte delle organizzazioni nei propri processi.

3 Servizi cloud come catalizzatori per l'adozione dell'AI

Nel seguente capitolo analizzeremo le applicazioni del cloud computing nell'adozione dell'AI. Ci concentreremo sui benefici e le sfide correlate a tale scelta tecnologica, mettendo in luce i vantaggi e i rischi di ciascuna delle architetture cloud analizzate in precedenza.

I tools e i servizi di intelligenza artificiale cloud-based condividono le principali caratteristiche generali dell'architettura cloud, sia in termini di vantaggi che svantaggi, ma implicano inoltre dei tratti unici che consentono alle organizzazioni di sfruttare l'AI nei propri processi in maniera semplificata.
Poiché lo sviluppo e l'addestramento di un modello di AI è un processo costoso e che richiede molto tempo, questi modelli sono diventati una forma di proprietà intellettuale di alto valore, quindi rappresentato il principale vantaggio per l'adozione di un servizio di AI rispetto ad uno concorrente.

Le caratteristiche principali di questi strumenti comprendono alcuni tratti ereditati dal cloud, l'astrazione della complessità, l'automazione e la personalizzazione, ma con alcune declinazioni specifiche per questa tipologia di servizi.

Caratteristiche ereditate dal cloud
Essendo i servizi AIaaS parte dell'ecosistema cloud, ereditano le caratteristiche tipiche di quest'ultimo, analizzate nel capitolo precedente. Di seguito si intende mostrare come questi tratti abbiano un impatto significativo anche in questa tipologia specifica di applicazioni.
Le caratteristiche che intendiamo analizzare sono: astrazione tra la risorsa sfruttata e l'infrastruttura sottostante, scalabilità e pagamento in base al consumo.

La scissione virtuale tra la risorsa sfruttata e l'infrastruttura sottostante ha ulte-

riore importanza in questa applicazione, infatti, in questo modo le stesse risorse possono essere condivide per differenti esecuzioni e processi, migliorando l'efficienza dell'utilizzo della potenza di calcolo. In tal senso, la scalabilità delle risorse ricopre un ruolo fondamentale, in quanto la possibilità di rilasciare ulteriori risorse in tempi brevi permette di gestire operazioni in parallelo e i picchi di carico. Poiché gli algoritmi di AI si basano sulla conoscenza estratta da una vasta mole di dati, la loro esecuzione richiede l'allocazione di risorse computazionali significative. Infatti, la scalabilità è vantaggiosa perché i requisiti di risorse variano molto a seconda delle operazioni eseguite, ad esempio, il training dei modelli può richiedere una potenza di calcolo molto più elevata rispetto all'esecuzione di quest'ultimo, ma per un breve periodo di tempo.
Con un modello di prezzo di pagamento in base al consumo, i clienti possono monitorare e controllare la spesa, infatti, i servizi AIaaS permettono di sfruttare le economie di scala, sia per i fornitori che per le organizzazioni. Gli utilizzatori non devono sostenere i costi iniziali per l'hardware, i costi di manutenzione e aggiornamento, potendo sfruttare i servizi di AI a basso costo.

Astrazione della complessità

La conoscenza tecnica e il know-how necessario alla realizzazione di applicazioni di AI proprietarie non sono facilmente accessibili alle organizzazioni, soprattutto alle PMI, a causa della ridotta disponibilità di figure professionali specializzate disponibili sul mercato del lavoro. Per questo motivo, i servizi AIaaS semplificano lo sfruttamento dell'AI attraverso l'astrazione della complessità.
Per astrazione della complessità si intende la possibilità di sfruttare tali strumenti con un approccio a *black box*, ovvero senza dover conoscere i dettagli e il preciso funzionamento dei processi sottostanti così come l'hardware su cui vengono eseguiti.

Questa caratteristica permette agli utenti di poter ottenere dei ridotti time-to-market nell'integrazione dei servizi AIaaS nei propri processi. L'organizzazione

che intende sfruttare un modello di AI non è obbligata a partire da zero nella sua realizzazione, ma a seconda delle sue esigenze può selezionare un servizio con il grado di astrazione desiderato, valore inversamente proporzionale al grado di personalizzazione che intende attuare, e quindi dedicarsi esclusivamente alla configurazione e customizzazione del servizio e al suo sfruttamento.
Inoltre, l'abbattimento di queste barriere di ingresso in termini di know-how, genera una riduzione del grado di incertezza sull'efficienza e la qualità del risultato finale: i servizi AIaaS potendo essere sfruttati in tempi anche molto brevi, permettono alle organizzazioni di avere degli strumenti pronti all'uso per effettuare delle sperimentazioni e analisi di fattibilità sul processo che intendono realizzare, e successivamente virare verso soluzioni maggiormente personalizzate o proprietarie nel caso in cui il risultato dei test faccia emergere tale necessità.
L'astrazione riguarda anche la base dati utilizzata per l'addestramento del modello, ovvero il cliente finale non è costretto a raccogliere e fornire tali dati, ma può sfruttare direttamente modelli pre-addestrati forniti dai provider.

Riassumendo, il vantaggio dell'astrazione della complessità dei servizi AIaaS, trasferisce da cliente a fornitore l'onere e i costi: della gestione e manutenzione dell'hardware, dell'aggiornamento e l'eventuale addestramento del modello e dell'assunzione di personale specializzato.

Automazione

I servizi AIaaS implementano innovativi sistemi di automazione che consento di ottimizzare in maniera autonoma i modelli di AI, a seconda dell'utilizzo, e permettono di gestire senza la necessità di azioni manuali le risorse hardware e gestirne il carico e i guasti.
Utilizzando modelli di AI diviene fondamentale la loro ottimizzazione in termini di parametri del modello e dei dati su cui è stato allenato. Per questo motivo, i provider forniscono degli strumenti compresi nei servizi di AI per ottimizzarne queste caratteristiche in modo che l'utente non debba necessariamente agirvi in

manuale.
Inoltre, i fornitori prevedono anche dei processi automatici per gestire le risorse necessarie all'esecuzione di questi modelli, anche in caso di modelli customizzati, a seconda della tipologia di quest'ultimo che si esegue, adattando in maniera automatica l'hardware virtuale sottostante in base alle esigenze specifiche dell'algoritmo di AI. Gli automatismi nella gestione delle risorse riguardano anche la mitigazione dei disservizi in caso di guasto e le operazioni necessarie per la conseguente disaster recovery.

Il risultato dell'automazione di questi servizi è quello di migliorarne l'accuratezza, in termini di prestazioni ottenibili, oltre che il costo computazionale necessario per eseguirli, inoltre permette di assicurare un prodotto con un ridottissimo rischio di disservizio.

Personalizzazione
I servizi AIaaS non forniscono vantaggi solo alle organizzazioni con ridotte competenze tecniche e know-how ma forniscono anche gli strumenti adatti agli utenti con tali conoscenze per creare, modificare, allenare e configurare i propri algoritmi di AI. Queste possibilità di personalizzazione consentono di implementare i propri modelli di AI oppure customizzare e ottimizzare con i propri dati di training modelli preesistenti.

Tramite i servizi AIaaS i clienti possono sfruttare le opzioni di configurazione previste dal fornitore per ottimizzare il modello selezionato secondo le proprie esigenze e il caso d'uso di applicazione. Inoltre, effettuando operazioni di fine-tuning l'utente può adattare un modello pre-addestrato ma utilizzando i propri dati o una base dati raccolta specificamente per migliorare il modello a seconda dell'utilizzo.
Il **fine-tuning** è un processo secondo il quale un modello di AI pre-addestrato viene personalizzato e ottimizzato per svolgere specifiche operazioni. L'obiet-

tivo di questa operazione è quella di affinare le sue capacità su un dominio di informazioni circoscritte, inteso come un set di dati specifici relativi ai compiti che si intende far svolgere al modello adattato. Operativamente il fine-tuning si compone generalmente di due step:

- **Adattamento al dominio**: il pre-addestramento di un modello viene realizzato su di un set di dati altamente ampio, ed è più vasto tanto più si intende realizzare un modello general purpose. L'adattamento prevede invece l'utilizzo di un set di dati ristretto e specifico del dominio di interesse, in modo da migliorare le capacità del modello su questo set di informazioni.
- **Configurazione dei parametri**: durante il fine-tuning i parametri del modello vengono adattati per ottimizzarlo a seconda delle attività per cui si intende utilizzarlo e il set di dati con cui è stato adattato. Questo step viene svolto con varie iterazioni di test strutturati per identificare il valore corretto dei parametri che permette di ottenere le migliori prestazioni.

Questa operazione potrebbe però peggiorare alcune capacità del modello, al di fuori del dominio specifico su cui è stato adattato, per questo motivo a seguito del fine-tuning normalmente si restringe l'operatività del modello in modo che risolva le richieste solo se rientrano nel dominio specifico.

In alternativa al fine-tuning si può ottimizzare un modello di AI pre-addestrato tramite un processo detto **grounding**. È una metodologia che prevede l'estensione della conoscenza di un *LLM* (Large Language Model) con informazioni specifiche del caso d'uso, non disponibili come parte dell'addestramento del modello stesso. Il vantaggio del grounding è quello di non dover effettuare dispendiosi addestramenti ulteriori del modello, ma è sufficiente fornire a questo una struttura con un punto di accesso per reperire tali informazioni. Si tratta quindi di una soluzione molto più economica e rapida rispetto al fine-tuning per personalizzare le informazioni di un modello di AI.

Oltre all'adattamento di un modello preesistente, i clienti dei servizi AIaaS hanno la possibilità di realizzare i propri modelli proprietari, sfruttando anche componenti fornite dal provider o da terze parti. In quest'ottica, i servizi possono offrire una piattaforma di sviluppo e di testing dei modelli, in modo da semplificarne l'implementazione. Infine, il provider si occupa di fornire, in maniera trasparente e scalabile le risorse necessarie alle fasi di addestramento, testing ed esecuzione del modello.
A seguito dello sviluppo del proprio algoritmo di AI, esso può essere distribuito direttamente tramite i servizi stessi, in modo da poter essere integrato nei propri processi beneficiando comunque dei vantaggi delle architetture cloud.

Le possibilità di personalizzazione dei servizi AI cloud-based permettono anche agli utenti più esperti di realizzare o customizzare i modelli di AI in modo da poter implementare la soluzione più adatta alle proprie esigenze ma senza dover rinunciare a tutti i benefici delle soluzioni cloud.

3.1 Vantaggi di ciascuna architettura nell'ambito dello sviluppo di soluzioni di AI

Come mostrato nel capitolo introduttivo, i servizi di AI realizzati in cloud prevedono dei modelli di erogazione che coincidono con i classici modelli di adozione del cloud. In altre parole, possiamo ragionare, anche nel caso di questa tipologia di strumenti in termini di modelli SaaS, PaaS e IaaS.

Software as a Service

La tipologia più diffusa è, anche in questa categoria, quella dei servizi software di AI, correlabili al modello SaaS. Si tratta di applicazioni pronte all'uso o blocchi di costruzione integrabili nei propri processi. Tali strumenti offrono la possibilità di utilizzare modelli di AI o metodi di apprendimento automatico pre-addestrati, rimuovendo l'onere per l'addestramento e configurazioni dei modelli da parte

degli utenti finali. In questo modo i provider possono offrire i propri modelli o i modelli realizzati da altri attori del mercato direttamente agli utenti finali. Possiamo dividere i servizi di AI con metodo di erogazione SaaS in due macro-gruppi:

- **Tool di AI**: si tratta di applicazioni di AI complete, che vengono utilizzate per la risoluzione di specifici compiti. Alcuni degli esempi attualmente più diffusi di tool di AI sono:
  - *Grammarly*[18]: supporta gli utenti nella correzione e nel miglioramento dei testi
  - *Midjourney*[19]: permette la generazione di immagini sulla base di una descrizione fornita
  - *GitHub Copilot*[20]: è uno strumento realizzato da GitHub e OpenAI per supportare gli utenti nelle attività di sviluppo all'interno di un IDE
- **Applicazioni integrabili nei processi**: si tratta di servizi erogati con metodologia SaaS ma non direttamente utilizzabili, in quanto richiedono l'integrazione in un applicativo o in un processo esistente; quindi, coincideranno con un blocco di un processo o di un tool più ampio. Tra gli esempi più noti citiamo due soluzioni realizzate da Microsoft:
  - *AI Builder*[21]: si tratta di una suite di soluzione di AI appartenente alla piattaforma Power Platform, realizzata sempre dalla casa di Cupertino. AI Builder comprende un set di processi integrabili in altri servizi SaaS presenti nel pacchetto Power Platform. Offre un'ampia gamma di modelli di AI pre-addestrati integrabili nei propri processi per risolvere specifici compiti come la categorizzazione di un testo o analisi del sentiment di una comunicazione.
  - *Dynamics 365 Copilot*[22]: si tratta di un Chat Bot basato sui modelli GPT realizzati da OpenAI e integrabile nelle applicazioni basate sul servizio Dynamics 365 (offerto da Microsoft). Copilot offre un certo

grado di personalizzazione e della conoscenza del modello. È possibile ottimizzare lo strumento tramite i propri dati in modo che sia in grado di interpretare richieste specifiche oppure può essere utilizzato per la risoluzione di altri compiti come la scrittura di e-mail o il riassunto dei testi.

Il principale vantaggio di questo tipo di servizi è la ridotta conoscenza tecnica specifica necessaria al loro sfruttamento, anche organizzazione con un limitato know-how sulle tematiche di intelligenza artificiale possono utilizzarli o integrarli nei propri processi. All'opposto, offrono limitate o nulle possibilità di personalizzazione, quindi i clienti devono accettare le limitazioni e le funzionalità offerte dal provider, senza potervi agire.

Platform as a Service

La metodologia PaaS prevede servizi che comprendono una piattaforma di sviluppo dedicata alla creazione delle proprie soluzioni di AI. Gli sviluppatori sono supportati nella realizzazione della propria soluzione o nella personalizzazione di un modello pre-addestrato. A seguito dell'attività di sviluppo, lo strumento realizzato è distribuito direttamente tramite il servizio PaaS, sotto forma di applicativo richiamabile on-demand e integrabile nei propri processi.

Tramite questi servizi è possibile effettuare operazioni di fine-tuning del modello in modo da adattarlo ai propri dati oppure è possibile agire sui parametri di configurazione o del modello per ottimizzarla a seconda delle proprie esigenze. Il servizio PaaS fornisce anche le risorse necessarie per eseguire queste attività e un ambiente utile alla fase di testing del prodotto realizzato.

Si tratta di soluzioni adatte alle organizzazioni che intendono adottare uno strumento, almeno in parte, personalizzato secondo il proprio caso d'uso ma, grazie alla piattaforma di supporto con una semplificazione le attività di sviluppo, non

necessitando figure con una specifica conoscenza sul funzionamento del modello di AI di partenza.

Infrastructure as a Service

I servizi di AI erogati con metodologia IaaS si riferiscono all'offerta di una infrastruttura virtuale atta a fornire all'utente la potenza computazionale necessaria per creare, addestrare e testare la propria soluzione di AI, oltre che la capacità di rete e lo spazio di storage dei dati. Si differenziano dai classici servizi IaaS in quanto prevedono delle risorse ottimizzate per la realizzazione di questi prodotti, ovvero hardware specializzato nell'esecuzione di software di AI come GPU per l'addestramento, maggiormente performanti delle classiche CPU in questo ambito, o l'accesso a database non relazionali *NoSQL* più adatti alla memorizzazione di dati non strutturati utili all'addestramento.

Il principale vantaggio di questi servizi è la scalabilità delle risorse e il conseguente pagamento in base al consumo. L'addestramento e il testing e l'esecuzione dei modelli di AI è un'attività con un costo computazionale molto elevato; in particolare le fasi di addestramento e testing richiedono un'infrastruttura hardware molto performante ma per un tempo relativamente breve, completate queste fasi l'esecuzione del modello potrebbe necessitare invece di una capacità minore. Per questo motivo, l'installazione delle risorse utili alla realizzazione del proprio modello in modalità on premise risulterebbe poco vantaggioso per un'organizzazione. La stessa *OpenAI* sfrutta i servizi IaaS offerti dalla piattaforma cloud *Microsoft Azure* per l'addestramento e l'esecuzione dei propri processi[12].

3.2 Sfide legate all'utilizzo di servizi cloud per l'adozione dell'AI

L'adozione dell'AI tramite i service cloud può presentare diverse sfide: le principali riguardano i rischi ereditati dallo sfruttamento del cloud in generale, ma è essenziale notare come in questa situazione sia di maggiore rilevanza la privacy dei dati e i rischi connessi all'utilizzo di modelli pre-addestrati.
Analizzeremo di seguito le tre principali sfide: sicurezza dei dati, vendor lock-in e utilizzo di modelli pre-addestrati.

Sicurezza dei dati

I servizi AIaaS ereditano i rischi correlati all'utilizzo del cloud in termini di sicurezza e privacy dei dati. In particolare, non avendo accesso diretto alle risorse, per l'utente finale è impossibile attuare delle strategie per mitigare questi rischi, infatti, i dati memorizzati in cloud sono in gestione del provider che si deve occupare della loro manutenzione.
Per questo motivo possono verificarsi delle situazioni di violazione dei dati, pensiamo ad esempio a una base dati utilizzata dagli utenti per delle attività di fine-tuning, o anche la perdita dei propri modelli creati tramite risorse PaaS o IaaS.

Vendor lock-in

Nelle applicazioni di AI ancora di più che nell'utilizzo in generale, lo sfruttamento del cloud si porta dietro un importante svantaggio correlato al cosiddetto vendor lock-in ovvero la dipendenza dal fornitore del servizio. Se nei servizi cloud classici questo rischio era dovuto ai costi e al tempo richiesto ad un'organizzazioni per migrare le proprie applicazioni da un provider ad un suo competitor, nel caso dell'AI la situazione si complica ulteriormente, in quanto, molti provider offrono le proprie soluzioni di AI comprese di modelli proprietari e addestrati su dati raccolti dallo stesso provider.

Perciò, nel caso in cui un'organizzazione intendesse modificare il proprio fornitore di servizi cloud non dovrà solo sostenere i costi correlati alla migrazione del servizio ma dovrà anche adattare i processi ad esso connessi ai modelli di AI offerti dal provider verso cui si sta migrando.

Modelli pre-addestrati

Le organizzazioni che intendono sfruttare modelli pre-addestrati proprietari dei provider spesso potrebbero non essere a conoscenza di molti dettagli relativamente a come è stato sviluppato e addestrato il modello.
In particolare, le aziende potrebbero non avere informazioni sui dati con cui il modello è stato addestrato, generando elevati rischi in termini della qualità del risultato e dovendo quindi sottostare alla possibilità di errore del modello. Per questo motivo, è importante effettuare dei test approfonditi anche nel caso in cui si intenda utilizzare un modello fatto e finito, in modo da verificarne l'accuratezza per le attività per cui si intende applicarlo.

In secondo luogo, a causa dell'elevata mole di dati necessaria ad addestrare e aggiornare un modello di AI, i fornitori di quest'ultimo potrebbero raccogliere e memorizzare i dati utilizzati dal cliente finale nell'esecuzione del modello, per poterli sfruttare come fonte per future versioni. L'utilizzo di questi dati da parte del provider potrebbe comportare degli importanti rischi in termini di privacy per le organizzazioni: nel caso in cui il provider sfruttasse i propri dati per future versioni del modello vi è il rischio che tali informazioni vengano indirettamente rese accessibili al pubblico.
Perciò, è fondamentale per le aziende informarsi sulle politiche di conservazione dei dati che transitano nei servizi di AI, e come questi dati possono essere sfruttati in futuro.

4 Soluzione proposta

Nel seguente capitolo è riportata una panoramica del contesto e dei presupposti che hanno condotto al presente studio e alla progettazione e realizzazioni del caso d'uso che sarà analizzato nei seguenti capitoli, basandosi sulle esigenze degli stakeholder e sulle osservazioni emerse dalle analisi precedenti. In particolare, si fornirà una visione d'insieme della società Cluster Reply presso la quale è stato svolto questo lavoro di tesi e delle premesse tecnologiche e funzionali su cui si è fondata la realizzazione di tale progetto.

4.1 Cluster Reply

Cluster Reply[23] è una società di consulenza e system integration che si focalizza sullo sviluppo di soluzioni basate su tecnologie Microsoft. Fa parte del gruppo **Reply**[24], realtà fondata nel 1996 a Torino e operante secondo un modello a rete di aziende altamente specializzate. Attualmente Reply opera a livello globale: dall'Europa ha espanso il suo operato in America, Asia e Oceania, con la presenza di più di 50 sedi sul territorio che contano oltre 12500 addetti a livello mondiale.

Cluster Reply è un player di riferimento del mercato italiano per quanto concerne lo sfruttamento delle tecnologie Microsoft, infatti, con i suoi 25 anni di esperienza e collaborazione con questo provider offre soluzione innovative ai suoi clienti in diversi settori industriali. In particolare, nel 2020, Cluster Reply si è aggiudicata il riconoscimento di *Microsoft Partner of the Year* per l'Italia, il quale premia il partner che a livello nazionale, nel corso dell'anno, si è distinto nella propria offerta di soluzioni innovative basate sulle tecnologie Microsoft.

La società opera in tutti i principali settori industriali, proponendo soluzioni atte ad innovare i processi aziendali, progettarne l'architettura e realizzare applicazioni dedicate alle necessità dei clienti. Le competenze si concentrano in

particolare nelle aree di *Modern Work*, *Security*, *Apps & Infrastructure*, *Digital Innovation*, *Data & AI* e *Business Applications*.

La company è suddivisa in sotto entità, dette virtual company, sulla base del proprio core business, questi enti sono a loro volta ripartiti in business units a seconda della industry di riferimento su cui operano. In particolare, il presente progetto è stato realizzato presso **Cluster DCX**, divisione della società che si occupa della realizzazione di *Business Application*, sviluppate sfruttando principalmente le tecnologie cloud offerte da Microsoft quali Power Platform, Dynamics 365 e Azure. Inoltre, i clienti oggetto di questo progetto si collocano tutti nella industry *automotive*, settore di riferimento della business unit presso il quale è stato realizzato.
All'interno della divisione Cluster DCX vengono realizzate soluzioni altamente personalizzate: l'offerta non si focalizza esclusivamente sull'attività di sviluppo ma copre tutte le fasi del progetto, infatti, la società si occupa anche della progettazione dello strumento sulla base dei requisiti e dell'infrastruttura, generalmente cloud, su cui lo esso sarà eseguito. La realizzazione di questi strumenti è corredata anche dallo sviluppo di processi correlati, come mobile app, analisi dei dati e artificial intelligence.

Le applicazioni sviluppate coprono moltissimi casi d'uso, ma possiamo riassumere i principali:

- Customer Relationship Management (CRM)
- Customer Service Applications
- Issue Management Applications
- Reporting Tools

4.2 Fasi del progetto

Le soluzioni analizzate nei capitoli successivi, nate a seguito di un progetto di ricerca interno e a specifiche richieste da parte di clienti del gruppo, si basano sulla progettazione e realizzazione di soluzioni di AI applicabili a diverse tipologie di business applications, in modo da automatizzarne i processi e supportare gli utenti di queste piattaforme nelle loro attività.
Il progetto nella sua definizione generale ha previsto le seguenti fasi:

1. **Analisi dell'offerta di servizi AI cloud-based** offerti dalla piattaforma Microsoft Azure e dai competitors: al capitolo 2 è stata fornita una sintetizzazione di tale analisi, in ottica da mettere in luce i player del mercato analizzati. Tutte le fasi successive del progetto si sono concentrate esclusivamente sulla piattaforma Azure con l'ottica di prendere in esame i competitors solo nel caso in cui gli strumenti di tale servizio non fossero sufficienti o adatti a realizzare le soluzioni ipotizzate.

2. **Redazione di vari casi d'uso**: sulla base di proposte interne o relativi a richieste di clienti della società, le proposte realizzate sono state suddivise in due macro-gruppi:

   - *Chatbot*: inerenti tutti quei processi che prevedevano un'interazione in linguaggio naturale da parte dell'utente finale
   - *Automazione processi*: comprende tutte quelle applicazioni dell'AI per automatizzare alcuni processi altrimenti svolti in manuale, come compilazione di form, categorizzazioni di comunicazioni ricevute o identificazioni di similarità tra diversi set di dati.

3. **Progettazione della soluzione**: in primo luogo effettuando un'analisi di fattibilità, e successivamente identificando i servizi della piattaforma Azure utili a realizzare la soluzione. La scelta dell'infrastruttura cloud da sfruttare molto spesso non è stata univoca, in diversi casi che analizzeremo di seguito, sono state proposte diverse architetture, che si differenziavano in termini di servizio o metodologia di erogazione di quest'ultimo. Una volta

identificate le risorse si è tentato di effettuare un'analisi delle attività di sviluppo necessarie per realizzare inizialmente un Proof of Concept (PoC) e successivamente la soluzione completa e pubblicabile.

4. **Analisi dei costi e dei tempi di realizzazioni**, suddividendoli in:

   - Costi dell'infrastruttura cloud scelta
   - Tempi di realizzazione
   - Costi di sviluppo

   A seguito di tale analisi è stato possibile identificare le soluzioni con una delivery più rapida e un minore impatto in termini di costi.

5. **Realizzazione del PoC**: è stata effettuata una scelta dei casi d'uso e conseguenti infrastrutture cloud migliori in termini di applicabilità, costi e tempi di realizzazione. Le soluzioni così identificate sono state sviluppate sotto forma di PoC.

6. **Esecuzione di test** approfonditi sulla qualità della soluzione sviluppata, eventualmente insieme al cliente, in modo da valutare i vantaggi di ciò che era stato realizzato.

7. **Realizzazione delle soluzioni** per alcuni clienti, integrate con business applications esistenti: al termine della fase di testing, alcune soluzioni sono risultate altamente vantaggiose e stabili e quindi sono state realizzate per alcuni clienti e risultano quindi attualmente utilizzate in produzione.

Le fasi indicate sono state svolte anche in parallelo e da parte di diversi team dell'azienda, inoltre, è importante notare che il progetto, da un punto di vista generale non è terminato, attualmente si stanno ancora portando avanti queste fasi in diversi casi d'uso o si stanno identificando nuovi requisiti in cui poter applicare processi di intelligenza artificiale.

4.3 Confronto con gli stakeholders

Le conclusioni dell'analisi delle offerte di mercato riportate in precedenza e l'analisi sull'applicabilità di queste tecnologie hanno dimostrato una serie di possibilità di innovazione per integrare strumenti di AI nei processi delle organizzazioni. Queste analisi hanno costituito la primissima fase del progetto qui riportato, a cui in seguito è stato effettuato un confronto approfondito con una serie di stakeholders, interni ed esterni all'azienda, in modo da comprenderne le esigenze e le opportunità di realizzazione.

È stato quindi possibile ottenere il punto di vista del management di Cluster Reply, per valutare le opportunità che le tecnologie di AI potevano ricoprire nell'offering di quest'ultima e di diversi attori delle organizzazioni già clienti, tra cui referenti ICT, referenti business, utenti finali e management, in modo da comprenderne le richieste e l'interesse sulla tematica. Le aziende coinvolte in questa fase sono state principalmente due, appartenenti al settore automotive, che utilizzano al loro interno un'ampia varietà di applicazioni cloud realizzate da Cluster Reply.
Sono stati effettuati degli incontri diretti con questi attori, in modo da valutarne le richieste e l'interesse generale o applicato a specifici tool utilizzati. Questi confronti hanno permesso di raccogliere una serie di spunti legati agli ambiti di maggior interesse in cui poter applicare processi di AI, in particolare alle caratteristiche ritenute essenziali, ai casi d'uso in cui sfruttarli, alle metodologie di integrazione con l'attuale ecosistema software, e alle esigenze in termini di costi e scalabilità.

Un ultimo confronto è stato svolto con vari referenti di Microsoft, figure di riferimento nell'ambito delle tecnologie AI cloud-based, tra cui: membri dei team di prodotto di vari servizi offerti, referenti tecnici per l'adozione della tecnologia e membri del reparto vendite. Grazie a questo scambio è stato possibile analizzare più a fondo l'offerta di Microsoft nell'ambito delle tecnologie di intelligenza artificiale e valutare in maniera approfondita i costi di tali servizi. Inoltre, è stato

possibile, tramite il supporto dei referenti tecnici dei diversi prodotti identificati, considerarne le possibilità di utilizzo e le modalità di sfruttamento, basandosi anche su quanto emerso dagli incontri precedenti con gli altri stakeholder.

I risultati di questi momenti di confronto hanno confermato le richieste del mercato e la possibilità per Cluster Reply di integrare servizi di AI cloud-based nell'infrastruttura delle organizzazioni di cui era già fornitore. In particolare, di centrale importanza sono stati i requisiti emersi dai clienti, che si costituivano principalmente di due tipologie di utilizzo, già citate nei paragrafi precedenti, ovvero chatbot e automazione, con un'importante predilezione per la prima. Infatti, la totalità degli attori esterni ha riportato un particolare interesse nell'integrare un chatbot, personalizzato nelle funzionalità e nella conoscenza ad esso disponibile, nella propria organizzazione e nei propri applicativi.

Dal punto di vista tecnologico, tutti gli attori, sia interni che esterni, hanno trasmesso l'interesse di adottare esclusivamente soluzioni cloud, in nessun caso è emersa l'esigenza di adottare soluzioni on premise, con un particolare focus sulla scalabilità dei servizi, un costo basato sull'utilizzo e un'adozione a fasi. Inoltre, i requisiti si focalizzavano principalmente sull'integrazione dell'AI nei propri processi, con un limitato interesse nella personalizzazione avanzata delle soluzioni ma piuttosto sull'adattare modelli esistenti nelle proprie soluzioni, in modo da ridurre i costi e i tempi di sviluppo ed eventualmente estenderli con ulteriori sviluppi.

La fase di confronto con gli stakeholder ha permesso quindi di definire alcune caratteristiche funzionali e tecniche delle soluzioni che saranno progettate e sviluppate. In primo luogo, la soluzione deve essere esclusivamente realizzata in cloud, deve risultare facilmente integrabile in un'infrastruttura preesistente, permettere un'adozione a fasi via via incrementali, e, di conseguenza, prevedere dei costi flessibili in base all'utilizzo. Come riportato in precedenza, è poi necessario che la soluzione realizzata possa essere inserita in contesti e soluzioni

esistenti, con un ridotto grado di personalizzazione necessaria, adattandone il funzionamento al caso d'uso specifico.
Per questi motivi, ci si è concentrati sui servizi di AI offerti dalla piattaforma Azure, raggruppati sotto la dicitura **Azure AI**, le cui caratteristiche soddisfano molte esigenze emerse.
Nel seguito della trattazione entreremo nel dettaglio di questa piattaforma, analizzando i singoli servizi offerti e i lori utilizzi, informazioni che risulteranno essenziali per comprendere la fase di progettazione delle soluzioni e il conseguente sviluppo.

5 Caso studio: Azure AI

Nel seguente capitolo entreremo nel dettaglio della piattaforma Azure AI utilizzata per il progetto presentato, in particolare sarà analizzata nell'interezza della sua offerta e delle sue principali caratteristiche e sarà fornito un approfondimento per ognuno dei servizi direttamente utilizzati nello sviluppo del caso d'uso riportato al termine dell'elaborato.

5.1 Caratteristiche

Azure AI è la piattaforma di servizi cloud di intelligenza artificiale offerti dalla multinazionale americana Microsoft. Si tratta di un portfolio di strumenti dedicati agli sviluppatori e ai system integrator per la creazione e distribuzioni di soluzioni innovative basate sull'AI potendo sfruttare modelli e infrastruttura gestiti dal provider.
Sempre all'interno dei servizi di AI realizzati da Microsoft, consideriamo parte della piattaforma Power Platform, che al suo interno offre degli strumenti dedicati all'integrazione dell'intelligenza artificiale. Questa seconda piattaforma è tecnicamente basata sulla prima, si tratta di una diversa nomenclatura dei servizi, infatti, la piattaforma Power Platform comprende tutti i prodotti offerti in metodologia SaaS, invece la piattaforma Azure quelli offerti secondo PaaS e IaaS.

I servizi sono stati sviluppati per integrarsi naturalmente nell'infrastruttura Azure e negli altri strumenti realizzati dalla casa di Cupertino. L'offerta prevede tutti i paradigmi di applicazione, l'ottica è quella di offrire servizi adatti a tutti i gradi di personalizzazione, perciò, la piattaforma prevede degli strumenti di AI per:

- Sfruttare modelli di AI o interi processi nelle proprie applicazioni
- Personalizzare modelli di AI e addestrarli tramite i propri dati
- Creare modelli di AI o machine learning ex-novo

Nella seguente figura[9] è riportata l'intera offerta di servizi di AI realizzata da Microsoft. Lo schema risulta così suddiviso:

- **Power Platform**: omprende al suo interno servizi di AI di tipo SaaS, sfruttabili tramite altri della piattaforma stessa.
- **Azure AI**: nomenclatura che comprende i servizi di AI presenti nella piattaforma Azure, suddivisi in tre macroaree:
  - *Azure Applied Services*: si tratta di servizi specializzati, offerti sotto forma di API, per scenari specifici che permettono di accelerare lo sviluppo dei processi e assicurano l'utilizzo di un modello ottimizzato.
  - *Azure Cognitive Services*: comprende una famiglia di API di AI personalizzabili per applicazioni generiche che permettono di sfruttare modelli pre-addestrati personalizzabili o ottimizzabili tramite i propri dati.
  - *Azure Machine Learning & Analytics*: comprendono piattaforme per lo sviluppo, il training e il rilascio dei propri modelli di AI, accedendo a funzionalità predefinite per semplificare le attività degli sviluppatori.

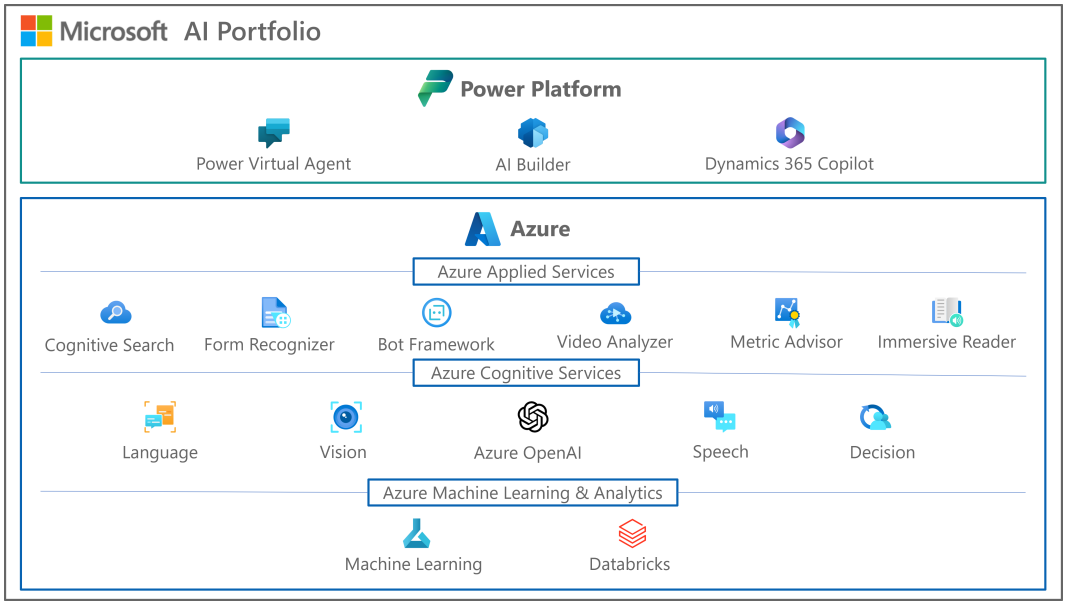

Figura 9: Microsoft AI Portfolio

Tutti i servizi possono essere gestiti tramite un interfaccia web-based, detta Azure Portal, che fornisce le funzionalità per crearli e configurarli. Inoltre, molti servizi presentano ulteriori interfacce dedicate, sempre accessibili sotto forma di applicazioni web, che fungono da ambienti di sviluppo, personalizzazione o testing delle funzioni specifiche da essi offerte.

5.2 Servizi utilizzati

Tratteremo di seguito i componenti della piattaforma di interesse di questa trattazione, in modo da fornirne una breve panoramica delle funzionalità, del modello di costo, della loro applicabilità e del paradigma cloud preponderante con cui sono erogati:

- **AI Builder**: si tratta di una suite di servizi di intelligenza artificiale, all'interno della Power Platform, integrabili nei propri processi mediante un approccio Low Code.

- **Dynamics 365 Copilot**: è una componente integrabile nelle web applications realizzate su piattaforma Dynamics 365 che permette di integrare in maniera nativa e con approccio No Code le funzionalità di un chatbot.
- **Cognitive Search**: si tratta di un servizio che permette di sfruttare funzionalità di ricerca avanzata e analisi del testo.
- **Form Recognizer**: consente di estrarre dati strutturati da documenti non strutturati.
- **Bot Framework**: si tratta di una piattaforma che supporta lo sviluppatore nella costruzione di interfacce di conversazione in linguaggio naturale.
- **Azure OpenAI**: consente di utilizzare i modelli sviluppati da OpenAI tramite un accesso con API.
- **Machine Learning**: si tratta di un servizio che permette di sviluppare, addestrare e distribuire modelli di machine learning e intelligenza artificiale.

Inoltre, è essenziale analizzare ulteriori servizi appartenenti alla piattaforma Azure ma non compresi nell'ambito dei servizi di AI, ma che sono stati utilizzati per realizzare i casi d'uso oggetto di questo progetto:

- **Data Factory**: consente di creare, pianificare e orchestrare trasformazioni e flussi di dati.
- **Data Lake**: fornisce una soluzione per l'archiviazione di una vasta gamma di tipologie di dati.
- **Function**: permette di rilasciare ed eseguire programmi realizzati in codice, detti appunto Function.
- **App Service**: fornisce un ambiente in grado di ospitare e rendere accessibili applicazioni web ed API.

Di seguito è schematizzata l'infrastruttura cloud dei servizi analizzati nei tre livelli di erogazione:

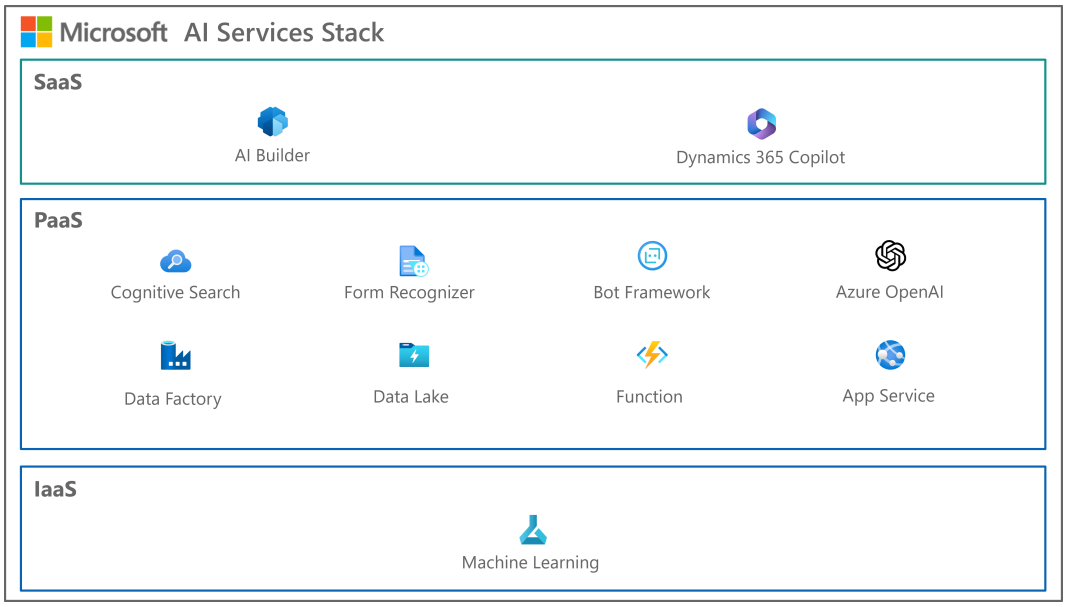

Figura 10: Microsoft AI Services stack

AI Builder

AI Builder è un servizio compreso nella piattaforma Microsoft Power Platform che fornisce modelli di intelligenza artificiale realizzati per ottimizzare i processi aziendali. Questo strumento consente alle organizzazioni di integrare processi di AI nelle proprie applicazioni realizzate sulla medesima piattaforma attraverso un approccio low code.

Tramite AI Builder è possibile creare modelli personalizzati in base alle proprie esigenze oppure utilizzare un modello predefinito pronto all'uso. Tra i principali modelli disponibili citiamo: classificazione di categorie, estrazione frasi chiave, analisi del sentiment di un testo e categorizzazione delle immagini. Il modello di interesse di questa trattazione è quello atto all'estrazione delle frasi chiave, il quale permette, dato un testo, di identificare i punti principali.
Questa componente può essere utilizzata mediante:

- **Power Automate**: si tratta di uno strumento low code che consente di creare dei processi automatizzati tramite la creazione di un flusso di blocchi predefiniti ma altamente configurabili. I flussi realizzati possono essere eseguiti in automatico, sulla base di specifici eventi o pianificati. Tali flussi possono eseguirsi nel contesto di molte applicazioni offerte da Microsoft come Dynamics 365 e Teams.

- **Power Apps**: è uno strumento low code per la creazione di web o mobile application. Tramite un approccio drag & drop è possibile strutturare la UI e mediante un linguaggio proprietario detto *Power Fx* o tramite dei flussi realizzati tramite Power Automate, è possibile gestire il comportamento della stessa e i processi che riguardano i dati.

Il costo di AI Builder è sottoforma di licenze che forniscono una specifica quantità di consumo dei modelli forniti. In altre parole, ogni licenza permette di avere a disposizione crediti di servizio per un certo periodo di tempo e ogni esecuzione di un modello predefinito o personalizzato è associato ad uno specifico valore di crediti consumati.
Una licenza ha un costo di 468,00 € al mese e fornisce 5000 crediti consumabili, i quali corrispondono, nel caso del modello di estrazione delle frasi chiave a 500000 esecuzioni su testi fino a 1000 caratteri.

Dynamics 365 Copilot

Dynamics 365 Copilot è uno strumento integrabile all'interno delle applicazioni realizzate tramite la piattaforma Dynamics 365. Questo strumento amplia le funzionalità dell'applicativo, fornendo agli utenti un'assistente integrato, tramite un chatbot che permette di supportare nelle attività di ricerca nella documentazione dell'applicativo o nella generazione di contenuti.

Questo prodotto è configurabile all'interno dell'ambiente a seconda delle funzionalità che si intende sfruttare, brevemente elencate di seguito:

- *Chatbot*: per ricercare contenuti all'interno del database dell'applicativo, interagendo in linguaggio naturale.
- *Creazione di contenuti*: permette di creare in automatico contenuti quali e-mail o risposte in chat, sulla base del contesto della conversazione e sul destinatario.
- *Riassunto dei contenuti*: fornisce la possibilità di creare riassunti dei contenuti del sistema in modo da velocizzare le attività degli utenti.

Dynamics 365 Copilot si basa sul servizio Azure OpenAI, analizzato di seguito, mettendo quindi a disposizione degli utenti un prodotto basato sui modelli GPT.

La licenza necessaria all'attivazione di questo prodotto rientra nelle licenze essenziali all'utilizzo della piattaforma Dynamics 365 e a seconda dei casi, potrebbe essere compresa o costituire un add-on da aggiungere.
L'organizzazione deve fornire una licenza per ogni utente a cui intende permettere l'utilizzo delle funzionalità offerte dal Copilot. Per esempio, la licenza *Customer Service Enterprise*, prevista per i sistemi *Dynamics Customer Service* per le realtà enterprise, ad un costo di 88,90 € per utente/mese, comprende attualmente anche il Copilot.

Azure OpenAI

Il servizio Azure OpenAI è il risultato della partnership tra Microsoft e OpenAI. Infatti, questo strumento permette l'accesso, mediante API, ai modelli di linguaggio creati da OpenAI. Questi modelli possono essere utilizzati nella loro distribuzione pubblica o possono essere personalizzati per eseguire specifiche operazioni o ottimizzati per uno specifico dominio tramite il fine tuning.

Microsoft offre uno strumento detto *Azure OpenAI Studio*, per gestire, testare e personalizzare le istanze di questi modelli.

Dopo aver creato una risorsa Azure OpenAI è necessario, proprio tramite questo strumento, distribuire un'istanza del modello che si intende utilizzare. A questo punto, è reso disponibile un endpoint con cui interagire per sfruttare il prodotto.

Risulta di fondamentale importanza la configurazione del **system message**, questo messaggio, inviato in ogni chiamata all'endpoint definisce il contesto della conversazione e il comportamento che deve tenere il modello nella generazione della risposta. Tramite il system message è anche possibile istruire il modello in modo da eseguire esclusivamente specifici compiti o obbligarlo a strutturare la risposta testuale in un determinato formato utile alle esigenze del processo.
La sezione detta *Playground*, presente all'interno della piattaforma Azure OpenAI Studio permette di testare a fondo le istanze create; tramite una UI a chat preconfigurata è possibile testare il modello rilasciato agendo sui parametri o sul system message in modo da identificare i valori che generano il maggior grado di ottimizzazione e il risultato ideale.

Il servizio elabora il testo suddividendolo in token, essi corrispondo a parole o a blocchi di caratteri: una parola brevi corrisponde ad un singolo token; invece, una di lunghezza maggiore sarà suddivisa in blocchi.
Il numero di token elaborati per ogni richiesta dipende dalla lunghezza dei parametri di input e output.

Azure OpenAI è un servizio *stateless*, ovvero un processo isolato che non consente la memorizzazione dei dati della conversazione corrente o di quelle passate, ogni interazione tramite le API è scollegata dalle precedenti. Per questo motivo, ogni esecuzione del modello prevede la chiamata al metodo esposto dall'API inviando tre informazioni:

- **System message**: viene inserito in ogni comunicazione e, grazie al comportamento stateless del servizio, può essere differente anche tra due chiamate successive, in modo da ottenere un diverso comportamento sulla base di richieste precedenti.

- **Prompt inserito dall'utente**: ogni esecuzione è avviata dall'utente nel momento in cui inserisce il proprio messaggio.
- **Stato della conversazione**: nella chiamata è inserito lo stato della conversazione, sottoforma di un certo numero di messaggi precedenti che si intende fornire al modello per comprendere nuovamente, di volta in volta, il contesto della conversazione. Tali dati testuali precedenti corrispondono sia ai prompt inseriti dall'utente nelle esecuzioni passate, ma anche le risposte generate dal modello sulla base di tali dati. Superiore sarà il numero di messaggi precedenti inviati migliore sarà il risultato ottenuto, ma all'opposto maggiore sarà il consumo di token.

Il risultato dell'esecuzione sarà un testo, generato dal modello, tramite questi tre dati forniti.

La principale differenza tra il servizio Azure OpenAI e i servizi offerti direttamente da OpenAI si colloca a livello dei vantaggi generati dallo sfruttamento della piattaforma Azure: utilizzare il servizio incorporato nella piattaforma semplifica l'integrazione con le altre risorse, inoltre Microsoft assicura la sicurezza e la privacy dei dati, permettendo anche la creazione di risorse all'interno di reti private.

Il costo della risorsa è correlato al numero di token consumati nelle esecuzioni e dal tipo di modello utilizzato e dal numero di parametri dello stesso.
Il modello *GPT-3.5-Turbo* a 16k parametri ha un costo di 0,003 € ogni 1000 token, invece la versione più recente dei modelli GPT, ovvero la *GPT-4* a 32k parametri equivale a 0,028 € ogni 1000 token.

Cognitive Search

Cognitive Search è un servizio di ricerca avanzata che fornisce agli sviluppatori l'infrastruttura e gli strumenti per configurare un ambiente di ricerca su contenuti eterogenei e le API utili ad eseguire operazioni di ricerca su quest'ultimo.

Scenari comuni possono essere la ricerca all'interno di documenti o di dati codificati su tabelle SQL.

Questo strumento mette a disposizione una serie di funzionalità per effettuare l'indicizzazione dei propri dati su uno o più indici di ricerca, in modalità full-text o tramite analisi lessicale. Una volta creato l'indice, tramite l'API esposta è possibile effettuare delle operazioni di ricerca, sotto forma di ricerca testuale, vettoriale, semantica, o fuzzy tramite l'esecuzione di chiamate ai metodi utilizzando una sintassi di query dedicata.
Possiamo quindi dividere il servizio in tre componenti:

- **Indexer**: componente atta alla trasformazione dei dati forniti in indici strutturati secondo la configurazione inserita, i dati forniti vengono elaborati in token e memorizzati in un indice. Il contenuto fornito può essere memorizzato in diverse modalità, dette origini dati: *Azure SQL Database*, *Azure Cosmos DB*, *Azure Data Lake* e *Azure Blob Storage*, in modo da permettere l'indicizzazione di dati di diverso tipo, strutturati e non strutturati. Tramite l'indicizzazione è possibile eseguire anche operazioni di arricchimento tramite AI, sfruttando altri servizi offerti dalla piattaforma, in modo da ottimizzare l'elaborazione dei dati o ampliarla con ulteriori informazioni, ad esempio l'arricchimento permette di:
  - *Analizzare le immagini*: tramite questa estensione è possibile indicizzare una descrizione delle immagini presenti nei documenti forniti, migliorando la qualità delle ricerche.
  - *Traduzione automatica*: questa estensione permette di tradurre i dati contestualmente all'indicizzazione, in modo da creare indici in ulteriori lingue su cui poter eseguire la ricerca avanzata.
  - *Rilevamento informazioni personali*: tramite questa estensione di analisi del testo è possibile riconoscere automaticamente le informazioni personali presenti nei dati da indicizzare e ignorarle o renderle non ricercabili in modo da rispettare vincoli di privacy.

- **Index**: componente per la memorizzazione dei dati trasformati secondo la modalità scelta. Sono formati da documenti di ricerca, che concettualmente corrispondono alla singola unità ricercabile, ogni unità è composta da campi, che rappresentano le informazioni su cui vengono eseguite le ricerche.

- **Query engine**: accessibile tramite i metodi dell'API esposta, permette l'esecuzione di query di ricerca complesse tramite un linguaggio di query dedicato. I risultati delle ricerche contengono un vettore di documenti di ricerca estratti secondo la query inserita e la logica di promozione dei risultati, ovvero, ogni risultato ottenuto dalla ricerca è associato ad un grado di confidenza del risultato, questo grado di confidenza è dovuto a tre fattori:

  - *Query eseguita*: a seconda della tipologia di query e dei parametri inseriti il servizio assegna un primo valore al grado di confidenza.
  - *Profili di scoring*: tramite questa funzione è possibile definire delle logiche di assegnazione personalizzate del grado di confidenza, tramite l'assegnazione di pesi specifici ai diversi campi dei documenti di ricerca.
  - *Semantic Search*: si tratta di un'estensione dell'esecuzione delle query che aggiunge un ulteriore livello di comprensione allo strumento, in modo da promuovere, incrementano il grado di confidenza, i dati semanticamente più rilevanti.

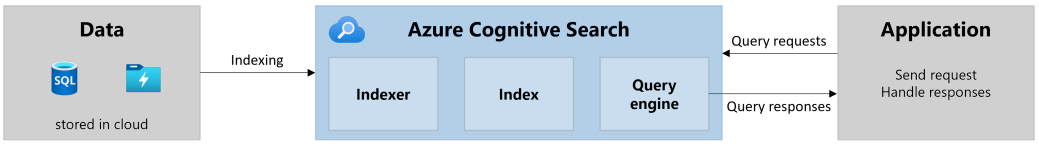

Figura 11: Schema componenti servizio *Azure Cognitive Search*

Microsoft ha messo a disposizione degli sviluppatori anche un ampio SDK per supportare e semplificare lo sviluppo di processi realizzati in codice per sfruttare questa risorsa, per esempio tramite le librerie presenti la scrittura della query è facilitata offrendo delle classi già costruite da dover solo valorizzare, invece di dover sviluppare la query da zero.

Il costo del servizio Azure Cognitive Search è strutturato in una serie di piani tariffari (o SKU), da selezionare al momento della creazione del servizio. Il livello selezionato determina:

- Numero massimo di indici consentiti
- Dimensione e capacità del servizio
- Costo del servizio ed eventuali costi incrementali per capacità aggiuntiva
- Possibilità o meno di utilizzare funzionalità aggiuntive come l'arricchimento tramite intelligenza artificiale o l'estensione semantic search

La capacità del servizio è misurata in termini di partizioni e repliche disponibili:

- *Partizione*: corrispondono allo spazio di archiviazione fisica degli indici e alla velocità delle operazioni di I/O, ad ogni partizione è quindi associato uno spazio di memorizzazione.
- *Repliche*: istanze del servizio, utilizzate per bilanciare il carico delle operazioni di ricerca. Ogni replica ospita una copia completa di ogni indice.
- *Unità di ricerca*: corrisponde alla componente minima di capacità del servizio, esiste un'unità per ogni partizione e replica, in altre parole essa contiene la sezione dell'indice presente nella partizione a cui è associata e la copia di tale sezione relativa alla replica a cui fa riferimento; perciò, il numero di unità di ricerca è dato dal prodotto del numero di partizioni e repliche selezionato.

I costi possono variare da 61,75 € a unità per il livello base, fino a 2352,53 € a unità per il livello premium, a cui corrispondono rispettivamente 2 GB e 2 TB di spazio di archiviazione per ogni partizione e un massimo di 3 e 36 unità di ricerca.

Form Recognizer

Form Recognizer è un servizio di AI che permette l'estrazione di dati strutturati da documenti non testuali. Tramite processi di riconoscimento ottico dei caratteri (*OCR*), permette di estrarre testo e dati da documenti di vari formati.

Il servizio offre un processo di estrazione automatica dei dati, estraendo l'intero contenuto del documento, diviso per pagine e righe, ma offre anche uno strumento per personalizzare il modello su specifici formati di documenti in modo da migliorarne l'accuratezza e strutturare l'output secondo le esigenze. In aggiunta, il modello permette di estrarre anche informazioni sul layout del documento.

L'esecuzione del modello avviene tramite una API esposta a seguito della configurazione della risorsa, i metodi che questa mette a disposizione permettono di eseguire tutte le funzionalità delle risorse, permettendo di elaborare documenti inviati in real time o memorizzati in un data lake e di ricevere il risultato del processo direttamente come risposta alla chiamata.
Anche in questo caso è disponibile un SDK, fornito per diversi linguaggi di programmazione, che permette di facilitare le attività di sviluppo per integrare la risorsa, in particolare i metodi dell'API esposta, nei propri processi.

Il costo del servizio è legato al consumo, inteso come numero di pagine elaborate, e al tipo di modello eseguito, differenziandosi quindi tra modelli predefiniti e personalizzati. In generale, il costo si attesta a 9,10 €/1000 pagine per i modelli predefiniti e a 45,49 € per i modelli personalizzati.

Bot Framework

Bot Framework è un servizio che prevede una raccolta di librerie e strumenti che consentono di creare, testare, distribuire e gestire interfacce di conversazione in linguaggio naturale o tramite la voce. Oltre a ciò, tramite questo prodotto è possibile implementare ulteriori processi, realizzati tramite un approccio low code o sviluppati via codice, da eseguire al verificarsi di specifiche situazioni durante la conversazione, come opzioni di reindirizzamento verso un agente umano.

Il risultato di quanto creato è poi integrato in un altro applicativo, che può essere una pagina web, una web app, o altri software offerti dalla stessa Microsoft come *Teams*. Le risorse necessarie alla gestione della conversazione e l'esecuzione del processo sono gestite sulla piattaforma Azure, l'applicazione accessibile all'utente finale presenta la UI e la comunicazione tra queste due risorse avviene tramite una API esposta dal servizio stesso.

Il costo di questo strumento è correlato al prezzo del servizio *App Service* su cui è rilasciato ed eseguito. Analizzeremo di seguito tale prodotto.

Machine Learning

Azure Machine Learning è un servizio per gestire il ciclo di vita dei progetti di modelli di machine learning o intelligenza artificiale. Permette di creare, addestrare e distribuire i propri modelli in modo efficiente e scalabile.

Il servizio prevede un ambiente di sviluppo, detto *Azure Machine Learning Studio*, per supportare le attività di realizzazione e training, sovrastante un'infrastruttura, detta *Azure Machine Learning Compute*, gestita in modalità PaaS o IaaS, a seconda delle proprie esigenze.I professionisti possono creare modelli custom o personalizzare modelli open source direttamente disponibili sulla piattaforma. Quest'ultima, semplifica le attività atte ad integrare il modello realizzato nelle applicazioni tramite strumenti che supportano il rilascio di API.

Tramite l'ambiente di sviluppo, l'utente può realizzare modelli e flussi di lavoro con un approccio low code, oppure sviluppare processi ex-novo, inoltre, la piattaforma supporta nell'esecuzione dei test ripetuti, nell'identificazione delle configurazioni che genera il massimo grado di ottimizzazione e nell'analisi delle metriche del modello rilasciato.

Il costo del servizio è correlato alla quantità di risorse riservate o consumate, infatti, tali risorse possono essere gestita secondo il modello PaaS, quindi con un pagamento in base al consumo, o secondo il modello IaaS quindi con la scelta delle risorse, intese come CPU, RAM e spazio di archiviazione, necessarie per le proprie esigenze. Il prezzo finale, in entrambi i casi, è molto variabile, per una risorsa con 2 vCPU e 8 GB il costo è di 79,70 €/mese ma questo valore può schizzare fino a 2550,34 €/mese per 64 vCPU e 256 GB di RAM.

Data Factory

Azure Data Factory è un servizio di data integration che consente di creare, pianificare e orchestrare flussi di dati da diverse fonti verso diverse destinazioni. È progettato per semplificare la gestione dei processi di estrazione, trasformazione e caricamento dei dati (*ETL*) da sorgenti non strutturate verso destinazioni strutturate.

Il servizio fornisce una piattaforma, detta *Azure Data Factory Studio*, che permette di gestire l'intero ciclo di vita degli *ETL*, con un approccio low code. Infatti, Microsoft mette a disposizione una serie di elementi predefiniti per strutturare il proprio flusso, per rilasciarlo e pianificarlo e per monitorarlo.

Il costo del servizio è correlato al numero di esecuzione pipeline rilasciate e alle risorse consumate da quest'ultime. Identificandosi quindi in un pagamento in base al consumo.

Data Lake

Azure Data Lake è un servizio per l'archiviazione progettato per la memorizzazione di grandi quantità di dati strutturati e non strutturati. Offre un ambiente di archiviazione scalabile che permette la gestione di qualsiasi dimensione o formato, inclusi file multimediali, documenti o dati di log.
Prevede nativamente funzionalità di sicurezza, automatismi di retention, versioning, meccaniscmi di ripristino e ridondanza dei dati.

Il costo è correlato allo spazio di archiviazioni e ai meccanismi di ridondanza e sicurezza selezionati.

Function

Azure Functions è un servizio serverless che consente di gestire processi realizzati lato codice senza doversi preoccupare della distribuzione e della gestione dell'infrastruttura. Inoltre, tramite questa piattaforma e l'SDK realizzato da Microsoft per molti linguaggi di programmazione, è semplificata l'integrazione dei propri processi con altri servizi dell'ecosistema Azure.
I processi distribuiti con questo servizio, detti *function*, possono essere realizzati nativamente con i linguaggi di programmazione più diffusi e sono forniti dei tool per diversi IDE atti a facilitare le attività di sviluppo e rilascio.

Il costo del servizio prevede un pagamento in base al consumo, correlato al tempo di calcolo sfruttato, oppure è possibile dedicare alcune risorse, intese come potenza di calcolo e RAM, tramite un *App Service Plan*, un piano che assicura la disponibilità di risorse sui cui eseguire diversi strumenti serverless.

App Service

Azure App Service è un servizio serverless per l'hosting di applicazioni web, API e processi back-end, basato su HTTP. È possibile realizzare, tramite i più diffusi

linguaggi di programmazione, i propri applicativi e rilasciarli su ambienti basati sia su Windows che su Linux.
Si tratta di un servizio analogo ad Azure Functions, con la differenza che in questo caso i prodotti sviluppati possono essere esposti tramite la rete.

Il costo della risorsa è analogo al caso precedente, ovvero con un pagamento in base al consumo oppure con l'acquisto di un *App Service Plan* tramite il quale vengono dedicate delle risorse al proprio servizio.

6 Caso d'uso: chatbot documentale

Nel seguente capitolo analizzeremo uno specifico caso d'uso, selezionato tra quelli che hanno superato tutte le fasi sopra indicate, ovvero che al momento della redazione del presente elaborato è stato rilasciato in produzione e quindi utilizzato da parte degli utenti finali, in quanto completo e stabile.

In particolare, esamineremo una soluzione di un chatbot documentale: un bot basato sui modelli GPT, esteso con i dati specifici dell'ente che intende utilizzarlo attraverso un processo di **grounding** per cui il modello possa estrarre conoscenza direttamente da una base dati documentale fornita e specifica per l'organizzazione.

Il modello GPT, sfruttato in versione GPT-3.5-Turbo, è un *Large Language Model* (LLM), progettato per essere utilizzato come motore di ragionamento generale per l'interpretazione e la gestione di testo. Questi modelli sono addestrati su un ampia varietà di informazioni, per dare loro una ampia comprensione del linguaggio e della manipolazione del testo, però non si tratta di database, ovvero non dovrebbero essere considerati dei depositi di conoscenza. Le informazioni che contengono, per quanto ampie, sono limitate: sono addestrati fino ad un certo tempo (ad esempio il modello indicato è stato addestrato con informazioni risalenti fino a settembre 2021) e non vengono aggiornati in maniera continua, inoltre i dati utilizzati sono pubblici e non specifici di alcun contesto.

Di conseguenza, tramite il grounding possiamo combinare le capacità di ragionamento dei LLM con le informazioni specifiche rilevanti per il contesto di interesse. Con questa soluzione possiamo evitare le dispendiose operazioni di training dei modelli, fornendo una soluzioni altamente scalabile e che permette un aggiornamento rapido e continuo dei contenuti. Tramite questo processo il modello può accedere, in tempo reale, al contenuto di interesse e rielaborarlo per soddisfare le richieste degli utenti.

Questa soluzione è stata studiata prendendo a riferimento tre diverse proposte infrastrutturali che prese in esame:

- *Soluzione SaaS*: una delle soluzioni proposte prevede un'architettura esclusivamente SaaS, integrabile nell'applicazioni realizzate tramite la piattaforma Dynamics 365.
- *Soluzione PaaS*: in questo caso la soluzione è stata realizzata sfruttando solamente servizi erogati secondo il modello PaaS offerto dall'infrastruttura Azure.
- *Soluzione PaaS + IaaS*: in quest'ultimo caso, l'architettura selezionate sfrutta alcuni servizi di tipo PaaS e altri di tipo IaaS.

L'intento sarà quello di riportate in breve le diverse fasi seguite e una trattazione più approfondita dei costi di realizzazione e il time-to-market, comparandoli tra le diverse architetture analizzate.

6.1 Perimetro delle proposta

Durante le prime fasi di progettazione, è stato necessario definire il perimetro della soluzione, in termini di funzionalità che saranno rese disponibili e componenti da utilizzare.
Questa attività risulta essenziale per la fase di progettazione della soluzione, in quanto permette di concentrarsi sulle funzionalità di maggiore interesse per offrire un prodotto completo e che soddisfa le esigenze degli stakeholders. In secondo luogo, consci del risultato che si intende ottenere è possibile realizzare un'infrastruttura cloud atta a soddisfare tali requisiti.

La soluzione proposta permette di integrare i modelli GPT con i dati delle organizzazioni, rappresentando un'estensione di tali modelli e un'evoluzione del

tool ChatGPT. Tramite un'infrastruttura decentralizzata sarà possibile integrare questo strumento negli applicativi aziendali, in modo da fornire un'interfaccia consistente con l'applicativo stesso, che consente agli utenti finali di interagire in linguaggio naturale per ricavare informazioni da varie fonti aziendali, sottoforma di risposte di testo strutturate. Il modello GPT sarà quindi in grado di consultare una serie di dati e documenti forniti, da cui poter estrarre i contenuti utili a soddisfare le richieste dell'utente finale.

L'integrazione tra gli applicativi e il chatbot deve essere scalabile, e non essere correlata a una specifica UI, in modo da poter essere adattabile a vari contesti. Per questo motivo, lo strumento deve permettere la segmentazione della conoscenza sulla base del punto di accesso o dell'utente che lo sta utilizzando. Infine, i contenuti forniti al modello devono poter essere facilmente aggiornati o estesi, in maniera autonoma da parte dell'utente finale.
Perciò, la soluzione realizzata prevede una struttura modulare, composta da diverse componenti, ognuna atta a gestire una fase del processo, in modo da renderla scalabile. L'accesso allo strumento avviene esclusivamente tramite API, di cui gli applicativi in cui è integrato ne implementano i metodi; invece, il cuore del processo rimane trasparente e gestisce le logiche di recupero, aggiornamento ed estensione di conoscenza. In altre parole, la soluzione deve comprendere le seguenti funzionalità:

- Integrare lo strumento in differenti sistemi in maniera indipendente.
- Gestire l'autenticazione dell'utente, permettendo quindi al modello di accedere esclusivamente alla conoscenza accessibile all'utilizzatore.
- Rispondere alle richieste degli utenti ricavando i contenuti dalla conoscenza fornita.
- Permettere di aggiungere o aggiornare documenti in autonomia.
- Processare i contenuti automaticamente, gestendo la segmentazione dei dati sulla base dei ruoli degli utenti e delle applicazioni da cui accedono.

6.2 Progettazione

La prima fase di progettazione della soluzione si è basata sulle conclusioni dell'analisi del mercato dei servizi cloud di AI e in particolare della piattaforma Azure AI, oltre che sui requisiti emersi dal confronto con gli stakeholder e definiti nella fase di modellazione del perimetro della proposta.
Si sono quindi identificate le caratteristiche tecniche di interesse e di cui tener conto nelle scelte infrastrutturali della soluzione:

- **Struttura scalabile e modulare**: la soluzione deve prevedere una struttura scalabile in termini di utenti che la utilizzano, inoltre deve essere modulare, per permettere di agire sui singoli componenti e non nella soluzione nella sua interezza, in modo da poter effettuare migliorie a fasi.

- **Processi standardizzati**: l'obiettivo era quello di realizzare uno strumento non pensato su misura per una singola organizzazione ma che presentasse processi standardizzati, adattabili alle esigenze del contesto specifico. Per questo motivo, nella progettazione si è tentato di realizzare uno standard di infrastruttura e processi, prevedendo meccanismi atti a adattarli nella maniera desiderata.

- **Personalizzazione**: la soluzione deve permettere diversi gradi di personalizzazione, potendo agire sulle componenti in maniera indipendente.

L'aspetto della standardizzazione merita particolare interesse, in quanto risulta essere una caratteristica essenziale della soluzione e del progetto, ovvero si tratta di una esigenza essenziale per la realizzazione di strumenti adattabili a diversi contesti e ambiti di applicazione senza dover richiedere sviluppi da zero ma potendo adattare il processo caso per caso. Inoltre, un alto grado di standardizzazione permette una riduzione dei costi, sia in termini di servizio, potendo sfruttare maggiormente le medesime risorse cloud, sia di sviluppo, semplificando alcune fasi nei diversi utilizzi e soprattutto di manutenzione, potendo gestire

soluzioni il più simili possibili tra loro.

Definiti i requisiti funzionali e tecnici si è passatati alla fase di progettazione della soluzione in termini infrastrutturali e di servizi cloud da utilizzare. Il prodotto è stato diviso nelle seguenti componenti:

- Componente per l'utilizzo di un LLM
- Gestione della base dati necessaria al processo di grounding
- Componenti per gestire e effettuare il processo di grounding
- Punto di accesso allo strumento per la sua integrazione in altre applicazioni

Tenendo a mente gli obiettivi, riportati sopra, di scalabilità, standardizzazione e personalizzazione, si sono strutturate diverse infrastrutture, con maggiori benefici in uno o più di questi ambiti e nei requisiti del prodotto. La progettazione ha portato quindi all'identificazione di tre infrastrutture, suddivisibili in parte secondo le tre metodologie di erogazione dei servizi cloud, riportate all'inizio di questo capitolo.
Andremo ora ad analizzare nel dettaglio queste strutture, evidenziandone gli aspetti preponderati secondo le caratteristiche di interesse del progetto.

Infrastruttura SaaS

Una delle infrastrutture progettate prevede esclusivamente l'utilizzo di servizi appartenenti alla tipologia SaaS. La soluzione ha come centro lo strumento Dynamics 365 Copilot.
La soluzione disegnata permette l'integrazione nativa in web application realizzate mediante Dynamics 365.

Il servizio Copilot offre una soluzione chatbot strutturata e tramite l'integrazione *out of the box* con le applicazioni Dynamics 365 è possibile interconnetterlo a quest'ultimo.

La gestione della base dati documentale è demandata a componenti direttamente presenti all'interno dei sistemi Dynamics 365: il processo di Knowledge Base Management. Tramite questa componente è possibile per gli utenti finali, aggiungere, modificare e rimuovere documenti non strutturati direttamente all'interno dello stesso applicativo. Questo processo è molto ampio nella sua forma base, permettendo un alto grado di possibilità di gestione per gli utilizzatori. La memorizzazione fisica dei dati avviene in un database non relazionale, compreso nel sistema Dynamics 365, detto Dataverse.

Tramite le funzioni di configurazioni è possibile strutturare il Copilot in modo da ricavare la conoscenza per rispondere alle richieste dalle fonti indicate. Lo strumento è stato quindi settato per recuperare le informazioni contenute esclusivamente nella base dati della Knowledge Base, ma rimane possibile fornire ulteriori fonti, anche esterne all'applicativo o all'organizzazione.

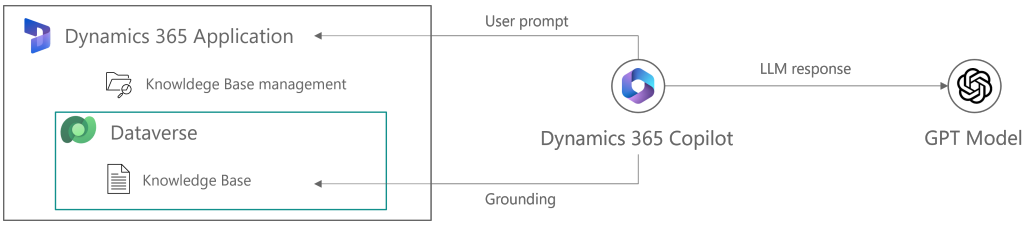

Figura 12: Schema infrastrutturale soluzione SaaS

Ulteriori opzioni di configurazione permettono di gestire altre funzionalità offerte dal Copilot, tra cui: creazione di risposte a conversazioni via e-mail e riassunto dei contenuti dell'applicazione. Inoltre, questo servizio permette di configurare tramite una funzionalità di ruoli di sicurezza quali utenti dell'applicativo possono utilizzarlo.

Infrastruttura PaaS

La seconda infrastruttura progettata e qui trattata prevede l'uso di diversi servizi offerti dalla piattaforma Azure AI, tutti di tipologia PaaS. Si tratta di una soluzione altamente più complessa rispetto al caso precedente, composta da varie componenti interconnesse tra loro, ognuna delle quali svolge una specifica funzione.

Il cuore della soluzione si identifica nel servizio Azure OpenAI, il quale mette a disposizione il modello GPT, sfruttabile tramite una API. Da questo nodo si diramano tre gruppi di componenti che gestiscono: l'accesso allo strumento realizzato tramite una API custom, il grounding per l'estensione della conoscenza tramite il servizio Azure Cognitive Search e un processo articolato su varie componenti per la gestione dei documenti utilizzati nel punto precedente.

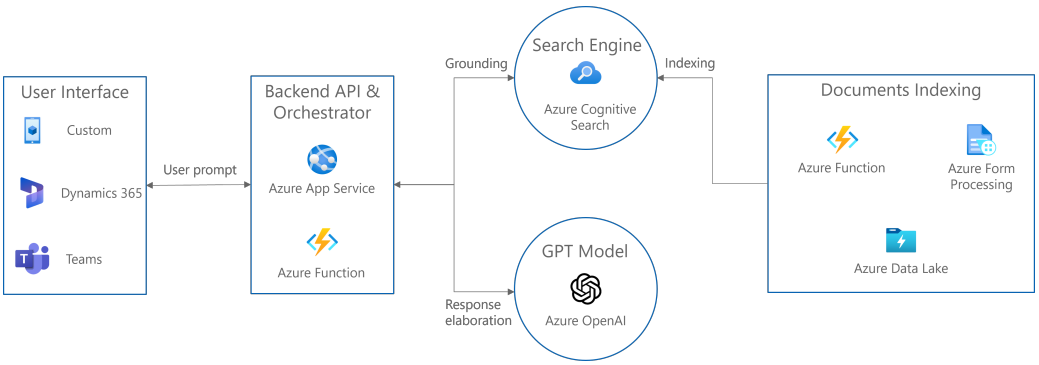

Figura 13: Schema infrastrutturale soluzione PaaS

Nella figura è schematizzata l'infrastruttura della soluzione, che possiamo suddividere in cinque aree:

- **User Interface**: l'interfaccia con cui l'utente finale interagisce con lo strumento. Tenendo a mente gli obiettivi del progetto questa componente può essere realizzata in molte tecnologie, in questo elaborato sarà mostrata una UI custom che permette di integrare lo strumento in applicazioni realizzate tramite il servizio Dynamics 365.

- **Backend API & Orchestrator**: comprende l'API, pubblicate su un servizio Azure App Service, che rappresenta il punto di accesso alle funzionalità del servizio e, strettamente connesso a questo, un orchestratore, realizzato mediante delle Azure Function che si occupano di gestire le richieste dell'utente.

- **GPT Model**: tramite il servizio Azure OpenAI è possibile sfruttare le potenzialità di un LLM, in questo caso è stato utilizzato il modello GPT-3.5-Turbo.

- **Search Engine**: mediante la risorsa Azure Cognitive Service, che si occupa di memorizzare un incide dei dati forniti alla soluzione, è possibile ricavare in tempo reale il contenuto di tali dati.

- **Documents Indexing**: corrisponde al processo di gestione della base dati fornita per il grounding. I documenti sono memorizzati in un Azure Data Lake, in cui possono essere inseriti o modificati in manuale, e processati dalle seguenti componenti:

  - *Azure Form Processing*: questo servizio estrae i dati strutturati dai documenti non strutturati.

  - *Azure Function*: gestisce il processamento dei documenti, occupandosi di elaborare i dati strutturati estratti dal servizio precedente e dividerli in singoli blocchi di informazioni per poi fornirle alla componente Azure Cognitive Search per effettuarne l'indicizzazione.

Analizzeremo in maniera dettagliata ogni area, le componenti che la compongono e le connessioni tra loro, ad eccezione dall'area di "Application UI/UX", non di interesse per questa trattazione, nel paragrafo di realizzazione della soluzione.

Infrastruttura IaaS

L'ultima infrastruttura progettata prevede l'utilizzo di diversi servizi della piattaforma Azure AI, sia di tipo PaaS che IaaS. In particolare, le parti centrali di questa soluzione sono analoghe al caso precedente, la differenza si identifica nell'area di *Documents Indexing*, nella quale a differenza di un processo realizzato tramite le risorse Form Processing e Function, si è ipotizzata una soluzione tramite la risorsa Azure Machine Learning.

Questo strumento permette di realizzare un processo di AI customizzato e allenarlo sui dati di interesse. L'ipotesi di questa soluzione è nata dall'assunto che, nel caso precedente, le componenti al minor grado di personalizzazione ma che potevano richiedere maggiori attività di ottimizzazione si collocavano proprio a livello del processamento dei documenti.

La risorsa Azure Machine Learning ci permette invece di sviluppare un processo personalizzato per questa soluzione, in modo da fornire maggiore libertà di ottimizzazione per la fase di indexing dei documenti.

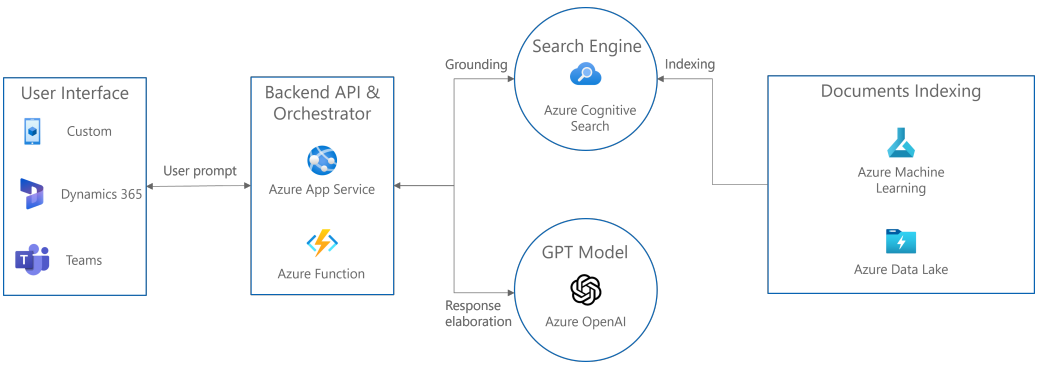

Figura 14: Schema infrastrutturale soluzione IaaS

Le restanti parti della struttura non sono variate rispetto al caso precedente, in quanto fornivano già il massimo grado di personalizzazione (sezione *Backend API & Orchestrator*) oppure l'alternativa di realizzare una soluzione custom risultava non applicabile (sezione *GPT Model* e *Search Engine*).

6.3 Realizzazione

La fase successiva del progetto ha riguardato la sua realizzazione. Di seguito sono riportati i diversi passi seguiti durante questa fase di sviluppo, raggruppati secondo le tre infrastrutture progettate al paragrafo precedente. Analizzeremo principalmente il processo sviluppato e le interazioni tra le componenti senza entrare esageratamente nei tecnicismi del lavoro, non essendo questo lo scopo dell'elaborato.

Soluzione SaaS

La realizzazione di questa soluzione ha previsto le seguenti fasi:

1. Installazione del servizio Dynamics 365 Copilot all'interno dell'applicazione Dynamics 365 e integrazione di questo strumento nell'interfaccia utente dell'applicativo.
2. Configurazione del servizio e delle sue funzionalità.
3. Configurazione del processo di Knowledge Base Management.
4. Gestione dei profili di sicurezza per l'utilizzo della risorsa.

È interessante notare come nessuna di queste fasi abbia previsto delle vere e proprie attività di sviluppo: si è trattato principalmente di attività di configurazione del prodotto e dell'integrazione in un'applicazione preesistente.

Tutte le attività svolte sono supportate dalla piattaforma Dynamics 365, che fornisce una interfaccia amministrativa che permette al system integrator di configurare le risorse in maniera guidata. Il risultato dell'integrazione di questi due strumenti è una sezione chatbot che viene presentata all'utente in alcune schermate dell'applicativo, in modo da fornire una esperienza integrata e consistente.

Soluzione PaaS

A differenza del caso precedente, la realizzazione della soluzione PaaS ha richieste molte più attività di sviluppo, prevedendo i seguenti steps:

1. Costruzione dell'infrastruttura dei servizi cloud.
2. Configurazione della risorsa Azure Cognitive Search, andando a definirne:
    - Le caratteristiche dell'index dei contenuti.
    - I parametri dell'indexer.
    - Le logiche di promozione dei risultati.
3. Configurazione della risorsa Azure OpenAI, effettuando il deploy del modello GPT-3.5-Turbo.
4. Configurazioni delle rimanenti risorse della struttura.
5. Sviluppo delle API, realizzando due metodi:
    - Metodo di autenticazione del client.
    - Metodo per la ricezione dei prompt dell'utente.
6. Sviluppo dell'orchestratore, realizzando un processo a fasi per gestire le richieste.
7. Sviluppo del processo di indicizzazione dei dati.
8. Configurazione della struttura del Data Lake per memorizzare i documenti e caricamento dei dati.
9. Sviluppo di una UI integrabile in una applicativo Dynamics 365.

La costruzione dell'infrastruttura avviene tramite gli strumenti della piattaforma Azure, il risultato di questa fase è stato la creazione di un resource group, contente le risorse necessarie.

Il servizio Azure Cognitive Search è stato configurando andando a creare un index e un indexer, per gestire l'indicizzazione di tutti i dati. La creazione di ulteriori indici è la modalità tramite la quale è possibile segmentare la conoscenza fornita allo strumento, in altre parole si potrebbe creare un indice per ogni gruppo di informazioni e ricercare i dati sull'indice corretto a seconda del contesto.
Attraverso la risorsa Azure OpenAI è stato effettuato il deploy del modello GPT-3.5-Turbo ed è stato quindi ottenuto l'endpoint per sfruttarlo.

Lo sviluppo dell'area di *API & Orchestrator* (relativa agli step 5 e 6) è stato effettuato tramite il linguaggio di programmazione *C#*.
I metodi previsti per l'API in questa soluzione sono stati due: un metodo necessario per l'autenticazione del client e il metodo principale per la ricezione dei prompt dell'utente. Però, è importante notare che, essendo stata realizzata questa API custom, questi metodi risultano molto scalabili, per esempio, in caso di particolari esigenze si potrebbero estendere per risolvere specifici compiti.

L'orchestratore, una volta ricevuta la richiesta da parte dell'API, sotto forma di prompt dell'utente, si occupa di gestire il processo per soddisfarla, seguendo tre steps:

1. Interrogando il servizio Azure OpenAI, genera una query ottimizzata alla ricerca delle informazioni utili all'interno del Search Engine, tramite il prompt dell'utente e lo storico della conversazione.

2. Esegue la query realizzata al passaggio precedente tramite il servizio esposto dal Search Engine e ricava quindi le informazioni utili sulla base della conversazione.

3. Interrogando nuovamente il servizio Azure OpenAI e fornendogli il prompt dell'utente, lo storico della conversazione, e le informazioni estratte al punto precedente in formato testuale, genera il testo di risposta che a sua volta viene tornato all'API.

È interessante notare il particolare utilizzo che è stato fatto del servizio Azure OpenAI, ovvero non è stato sfruttato esclusivamente per generare la risposta, ma anche per completare un compito assolutamente specifico e puntuale: generare una query per interrogare il servizio Azure Cognitive Search. Questo è stato possibile esclusivamente fornendo in questa prima interrogazione al servizio un system message che spiegasse, in linguaggio naturale, il risultato generale; ne forniamo qui un esempio per comprenderne meglio l'impatto:

"Below is a history of the conversation so far, and a new question asked by the user that needs to be answered by searching in a knowledge base about how to the use of company applications.
Generate a search query based on the conversation and the new question.
Do not include cited source filenames and document names in the search query terms. If the question is not in English, translate the question to English before generating the search query."

Di fondamentale importanza comprendere il grado di comprensione offerto dal modello GPT: tramite un messaggio molto semplice è possibile *istruirlo* in tempo reale a risolvere un task specifico, con una comprensione del contesto e dell'output desiderato.
Nella seconda interrogazione al servizio Azure OpenAI il system message varia (la scelta del system message da utilizzare viene fatta dall'orchestratore sulla base dello step corrente del processo), in una descrizione del comportamento necessario per generare la risposta, del tono da mantenere e soprattutto del fatto che le informazioni da utilizzare devono essere esclusivamente quelle fornite.

"Assistant helps the employees with their questions. Be brief in your answers.
Answer only with the facts listed in the list of sources below. If there isn't enough information below, say you don't know. Do not generate answers that don't use the sources below. If asking a clarifying question to the user would help, ask the question. For tabular information return it as an html table."

Le informazioni vengono indicizzate sull'indice tramite i processi raggruppati nell'area Documents Indexing, tramite un algoritmo, sviluppato in *Python*, e rilasciato sulla Azure Function che si occupa di gestire i documenti aggiunti, modificati e rimossi dal Data Lake, richiamare il servizio esposto dalla risorsa Azure Form Recognizer, per tradurre i dati non strutturati in informazioni strutturate e infine, tramite l'indexer, indicizzare tali informazioni.
Il servizio Azure Form Recognizer è stato utilizzato nella sua forma base, ovvero tramite il modello predefinito che permette di convertire dei documenti in dati testuali strutturati. L'algoritmo realizzato in *Python* si occupa di dividere le informazioni in singole unità ricercabili per poi essere memorizzate sull'indice.
Questo processo di preparazione dei dati ha richiesto particolare attenzione, in quanto divisioni differenti dei dati generano indici molto variabili, direttamente correlati alla qualità del risultato di ricerca. Per esempio, estendendo la dimensione dell'unità ricercabile si ottiene un indice meno dettagliato e quindi le query di ricerca estraggono informazioni più generiche rispetto alla richiesta.

L'ultimo passaggio, citato solo per completezza, ma non di interesse di questa trattazione, ha riguardato lo sviluppo di una interfaccia utente custom, realizzata mediante in *TypeScript* e il framework *React*, basato su *JavaScript*. Questa UI è stata quindi integrata in una applicazione Dynamics 365 esistente e messa in comunicazione con lo strumento tramite le API realizzate.

Soluzione IaaS

L'ultima soluzione analizzata ha sfruttato in parte quanto realizzato per il caso precedente, ma modificando la componente fondamentale di *Documents Indexing*. Come riportato in precedenza, l'algoritmo realizzato è un importante punto di attenzione della soluzione, per questo motivo si è ipotizzato di realizzare il medesimo processo ma in maniera maggiormente personalizzata, sviluppando un processo di AI tramite la risorsa Azure Machine Learning.
L'ottica di questa modifica era quella di poter realizzare un processo maggior-

mente articolato e preciso nella segmentazione dei dati in singole unità ricercabili. Anche la componente Form Recognizer in questa soluzione non è utilizzata, in alternativa si sfrutta uno strumento OCR mediante l'esecuzione di uno script *Python*.
Il processo, sviluppato tramite una struttura a blocchi, prevede una segmentazione ottimizzata dei dati, in modo da identificare le singole sezioni del contenuto, a sua volta dividendole in singole frasi. Lo scopo era quello di non spezzare frasi o sezioni in unità di ricerca differenti.

6.4 Analisi comparativa

Andremo ora a comparare le tre soluzioni realizzate, in termini di:

- **Costi di servizio**: con i dovuti assunti, in particolari in termini di utenti e consumo, saranno riportati i costi dei servizi per le tre infrastrutture analizzate.
- **Costi di realizzazione**: tramite una stima interna, sarà fornito un valore ipotizzato di costo di progetto per la realizzazione di queste soluzioni.
- **Costi di manutenzione**: analogamente al punto precedente, sarà fornito un valore ipotizzato dei costi di manutenzione di queste soluzioni, sull'orizzonte temporale del primo anno a seguito del go-live.
- **Time to market**: sarà riportata una stima del tempo che intercorre tra le fasi iniziali di un progetto di questo tipo, fino al momento del rilascio in produzione, per ognuna delle tre soluzioni.
- **Qualità del risultato**: tramite una serie di test standardizzati che sono stati svolti sulle soluzioni, in versione Proof of Concept, compareremo la qualità di quanto realizzato.

È fondamentale sottolineare gli assunti su cui si basano le seguenti analisi. In particolare, i costi di servizio sono calcolati sulla base dei prezzi pubblici dei servizi della piattaforma Azure, tramite o strumento Pricing Calculator[25], e tenendo a mente i seguenti parametri:

- Numero di utenti delle soluzioni: 100
- Il dimensionamento delle risorse è stato definito con un ampio scarto sulla base dei consumi generabili dal numero di utenti sopra riportati.
- Il costo delle risorse a consumo è stato definito ipotizzando un elevato utilizzo dello strumento da parte di ogni utente.

Inoltre, per quanto concerne i costi di realizzazione, di manutenzione e il time to market, i valori presentati si collocano in un'ampia forbice possibile, e stimati calcolando il costo in termini di giorni/uomo e il conseguente tempo di realizzazione, con uno scarto nei termini del 10%. Nella seguente tabella sono riportate le composizioni dei team ipotizzati per la realizzazione di ogni soluzione.

Tabella 2: Composizione dei team per ogni soluzione

Ruolo	**SaaS**	**PaaS**	**IaaS**
Project Manager	1	1	1
Business Analyst	1	1	1
Technical Leader	N/A	1	1
Developer	N/A	2	1
AI Developer	N/A	N/A	1

Costi di servizio

Mostriamo di seguito una tabella riepilogativa con i costi di tutte le componenti previste per ogni infrastruttura mostrata, i costi totali sono su base mensile.

Tabella 3: Costi di servizio soluzione SaaS

Soluzione SaaS	**Tier**	**Dettagli**	**Costo**
Dynamics 365 Copilot	Dynamics 365 Customer Service	100 licenze Enterprise	8890,00 €
		Totale	8890,00 €/mese

Tabella 4: Costi di servizio soluzione PaaS

Soluzione PaaS	**Tier**	**Dettagli**	**Costo**
Azure App Service Plan	P1V3	Windows	227,85 €
Azure OpenAI	GPT-3.5-Turbo	*16000 x 1000* tokens	29,55 €
Azure Cognitive Search	S1	2 unità di ricerca	453,01 €
Azure Form Recognizer	S0	*Pay as you go* - 5000 pagine	46,17 €
Azure Data Lake	Standard	1 TB - ZRS	14,04 €
		Totale	770,62 €/mese

Tabella 5: Costi di servizio soluzione IaaS

Soluzione IaaS	**Tier**	**Dettagli**	**Costo**
Azure App Service Plan	P1V3	Windows	227,85 €
Azure OpenAI	GPT-3.5-Turbo	*16000 x 1000* tokens	29,55 €
Azure Cognitive Search	S1	2 unità di ricerca	453,01 €
Azure Data Lake	Standard	1 TB - ZRS	14,04 €
Azure Machine Learning	D4V4	4 vCPUs e 16 GB RAM	183,36 €
		Totale	907,80 €/mese

Alcune note fondamentali, in primo luogo sembrerebbe che la soluzione SaaS sia nettamente la più costosa, ma è fondamentale comprendere che tale costo non riguarda esclusivamente il prodotto Copilot ma l'intera offerta dei servizi dell'applicativo Dynamics 365 Customer Service, per esempio, organizzazioni in cui questo strumento è già diffuso non sarebbero soggette ad ulteriori costi.

Inoltre, vale la pena chiarire il costo stimato della risorsa Azure OpenAI, come trattato in precedenza, esso è correlato al numero di token che vi transitano, in gruppi da 1000; in queste ipotesi, tenendo a mente un utilizzo da parte di 100 utenti è stato stimato un consumo di *16000 x 1000* tokens. Il valore 16000 è derivato dall'analisi degli accessi degli utenti ad un applicativo di Customer Service: in media gli utenti compiono due sessioni di utilizzo del sistema al giorno, si è quindi ipotizzato pari a quattro il numero di conversazioni con lo strumento per sessione, per i venti giorni lavorativi mensili.

La soluzione SaaS ha un costo direttamente prevedibile, rapportato esclusivamente al numero di utenti che accedono allo strumento, e per questo motivo i due valori sono direttamente proporzionali tra loro. Invece, per le altre soluzioni i costi non sono direttamente correlati al numero di utenti e nemmeno al consumo del servizio; infatti, l'infrastruttura realizzata permette di gestire carichi cospicui e scala su consumi molto più elevati. Per esempio, un raddoppio del carico non richiede particolari aumenti nei tier delle risorse e tramite i piani *Pay as you go*, i costi sono automaticamente rapportati all'utilizzo di quest'ultime.

Costi di realizzazione

Di seguito sono riassunti i costi di realizzazione per ogni soluzione, inoltre sono riportate le percentuali di questi che fanno riferimento ad ogni ruolo professionale all'interno del team.

Le stime sono fornite in un range ampio, in quanto sono altamente correlate alle attività necessarie per il contesto e il cliente per cui è realizzato.

Tabella 6: Costi di realizzazione per ogni soluzione

	SaaS	**PaaS**	**IaaS**
Costi di realizzazione	5 - 7,5 k€	40 - 50 k€	60 - 80 k€

Tabella 7: Percentuale dei costi inerenti i ruoli professionali dei team per ogni soluzione

Ruolo	**SaaS**	**PaaS**	**IaaS**
Project Manager	10%	5%	5%
Business Analyst	90%	10%	10%
Technical Leader	N/A	15%	15%
Developer	N/A	70%	40%
AI Developer	N/A	N/A	30%

Il costo è direttamente proporzionale al grado di personalizzazione della soluzione: maggiormente custom sono i processi sviluppati maggiori sono i costi ad esso connessi.

Interessante notare come, per la soluzione SaaS, non sia necessario nessun ruolo tecnico, questo permette un importante traslazione dell'onere di realizzazione tra ruoli tecnici e funzionali. Infatti, in questo caso il Business Analyst si occupa della realizzazione dell'intera soluzione, essendo richieste esclusivamente attività di configurazione, supportate dalla piattaforma Dynamics.

Spostandosi sulle architetture Azure delle soluzioni PaaS e IaaS, entrano in gioco il Technical Leader, per la gestione del team di sviluppo e che si occupa della realizzazione e configurazione dell'infrastruttura, e i Developer per le attività di sviluppo e testing. Infine, per la soluzione IaaS è fondamentale l'attività di un AI Developer, ovvero di un tecnico esporto sulla realizzazione di soluzioni di AI, figura maggiormente specializzata che genera un serio impatto sui costi.

Costi di manutenzione

I costi di manutenzione sono direttamente correlati ai costi di realizzazione: maggiore sono le attività di sviluppo e più articolata è l'infrastruttura, maggiori potrebbero essere gli interventi di manutenzione necessari.

Tabella 8: Costi di manutenzione per ogni soluzione

	SaaS	**PaaS**	**IaaS**
Costi di manutenzione	0 - 2 k€	7,5 - 10 k€	12,5 - 15 k€

La soluzione SaaS è ampiamente la più economica in termini di manutenzione: prevedendo sostanzialmente un'unica componente i rischi di malfunzionamenti infrastrutturali sono abbattuti, inoltre, non richiedendo attività di sviluppo, non si presenta il rischio di bug nei processi realizzati. In questo caso la manutenzione è correlata principalmente alle verifiche necessarie a seguito di ogni aggiornamento del servizio SaaS effettuato da provider.

Di contro, le soluzioni PaaS e IaaS presentano maggiori rischi, correlati a possibili problemi infrastrutturali dovuti all'elevato numero di componenti, oltre che a possibili bug che si verificano nel codice sviluppato.
Gli elevati costi di manutenzione, soprattutto nell'ipotesi IaaS, sono dovuti all'esigenza di figure specializzate anche per questa attività.

Time to market

Il time to market è correlato ai costi e tempi di realizzazione, ovviamente, maggiori sono le attività di sviluppo e personalizzazione richieste maggiori saranno tali tempi.
Inoltre, come è facile prevedere, team più numerosi generano minori time to market, in questa ipotesi si è tenuto conto di team dimensionati per massimizzare

gli sviluppi paralleli ma senza portare a una granularità della divisione della attività troppo diffusa, alla quale sono dovute attività di gestione e organizzazione ulteriori.

Tabella 9: Time to market per ogni soluzione

	SaaS	**PaaS**	**IaaS**
Time to market	2 – 3 settimane	12 – 15 settimane	14 – 18 settimane

La realizzazione della soluzione SaaS è sicuramente la più breve, però, è importante notare che per la messa in produzione di tale prodotto, è necessario avere già disponibile un applicativo Dynamics 365, prerequisito non scontato per tutte le organizzazioni.

Per le soluzioni PaaS e IaaS i tempi lievitano in maniera considerevole, ma non in modo direttamente proporzionale ai costi di realizzazione, in quanto le figure più specializzate presentano maggiori costi.
Inoltre, queste soluzioni necessitano di attività di testing assolutamente più impattanti rispetto alla prima casistica.

Qualità

Le verifiche sulla qualità della soluzione realizzata sono state eseguite mediante una serie di test standardizzati, applicati ai diversi strumenti sviluppati. In particolar modo, i test sono stati svolti mediante dei dati interni all'azienda Cluster Reply: sono stati identificati 10 documenti PDF da fornire allo strumento, per un totale di 124 pagine.
Sulla base dei dati forniti sono state dettagliate 10 domande e 5 conversazioni da fornire allo strumento, atte ad analizzare i comportamenti del sistema nel ricavare le informazioni necessarie dalla documentazione, richiedere ulteriori det-

tagli o a recuperare il contesto di una domanda dai precedenti messaggi della conversazione. Ogni domanda e conversazione è stata eseguita 5 volte per ogni soluzione, resettando il sistema a seguito di ogni test. Infatti, i test sono stati realizzati per valutare la qualità del processo di grounding delle diverse soluzioni e eseguiti nello stato di miglior ottimizzazione delle soluzioni.

Le risposte generate venivano poi valutate da un addetto in una scala da 0 a 3, dove 0 rappresenta *nessuna risposta fornita* e 3 *risposta eccellenze*, le risposte venivano penalizzate nel caso in cui le informazioni fornite venissero inventate o non risultassero corrette, oppure in caso di presenza di errori dal punto di vista linguistico, si è ritenuto sufficiente un risultato pari a 2.

La soluzione SaaS ha fornito una risposta sufficiente nel 53% dei casi, risultando migliore nelle domande singole rispetto che nelle conversazioni complesse. Però, è importante notare che, al momento di questi test il servizio Dynamics 365 Copilot si trovava ancora in uno stato di preview.
Le soluzioni PaaS e IaaS, nettamente migliori, hanno generato entrambe il medesimo tasso di risposte sufficienti, pari al 87% dei casi, risultando molto accurati anche nelle conversazioni complesse. Il caso IaaS è risultato solo più accurato in alcune risposte, fornendo maggiori dettagli correlati alla richiesta.

Nel seguente grafico sono riassunti i risultati, per i singoli test, per ogni soluzione, i valori mostrati corrispondono ad una media di tutte le iterazioni dello stesso test.

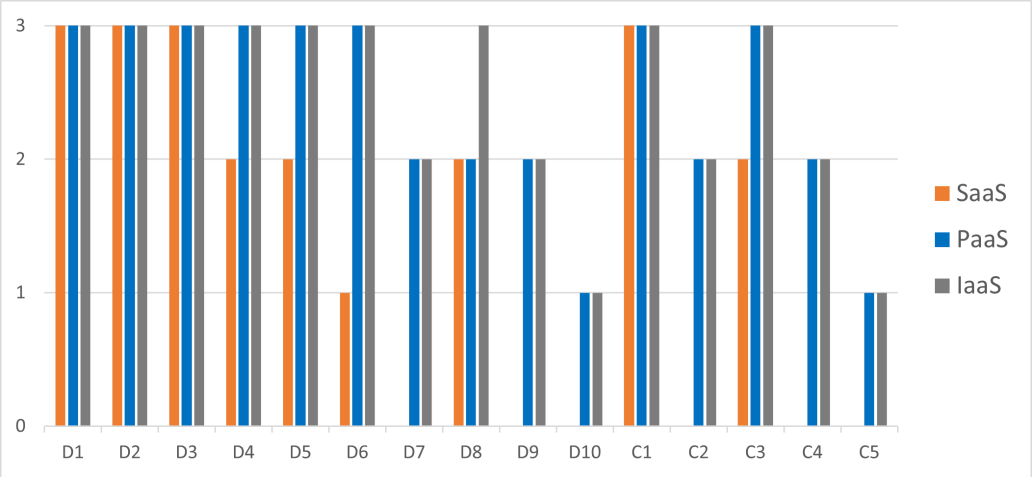

Figura 15: Risultati dell'analisi della qualità del risultato delle soluzioni realizzate

È interessante notare alcune tendenze nei risultati ricavati, in particolare, la soluzione SaaS tende maggiormente a non fornire nessuna risposta nel caso in cui non ricavi informazioni utili dai dati di grounding forniti rispetto alle altre due soluzioni, in questi casi invece, in alcuni test, sono state generate risposte inventate. Inoltre, le soluzioni PaaS e IaaS tendono ad inserire molti dettagli nelle risposte, alcune volte utilizzando informazioni non contestualizzate al contesto, in questo aspetto il modello SaaS tende ad essere più breve nelle risposte senza inserire informazioni al contorno non richieste.

Dai test effettuati è emerso che il processo di grounding custom realizzato per le soluzioni PaaS e IaaS risulti ampiamente migliore rispetto alla soluzione SaaS, permettendo di ritrovare risultati per il contesto della domande in un maggior numero di situazioni, però, all'opposto, il servizio SaaS risulta maggiormente ottimizzato nella costruzione della risposta, generandone di più lineari e precise quando ricava le informazioni utili, questo comportamento si presume dovuto ad una customizzazione effettuata dal provider sul modello GPT utilizzato, in modo da ottimizzarlo al meglio per il proprio servizio.
La realizzazione di un processo di indicizzazione altamente personalizzato, tramite un algoritmo di AI sviluppato per la soluzione IaaS, non genera il miglioramento sperato nella qualità del risultato, probabilmente il miglioramento

risulterebbe più evidente con un base dati fornita e indicizzata molto più ampia, ma allo stato attuale non sembrano giustificati i maggiori costi per realizzare tale soluzione.

Per trarre le conclusioni di quanto realizzato nel caso d'uso trattato, possiamo considerare tutte e tre le soluzioni realizzate ampiamente soddisfacenti rispetto ai requisiti forniti, inoltre, l'analisi comparativa delle strutture, ha permesso di ricavare dati fondamentali a comprendere i contesti di applicazione delle metodologie di erogazione dei servizi cloud.
Infatti, è stato compreso come in questa applicazione, la soluzione SaaS risulti una scelta eccellente per le situazioni in cui è richiesto un limitato tasso di personalizzazione e inoltre che richiede un time to market molto rapido. All'opposto, le soluzioni PaaS e IaaS risultano il miglior approccio per realizzare processi personalizzati e altamente ottimizzabili, correlati però a tempi e costi di sviluppo maggiori. Dato di assoluto interesse, e che ha contraddetto le ipotesi emerse in fase di progettazione, è stato il risultato dei test della soluzione IaaS che, pur avendo richiesto un maggiore investimento e figure maggiormente specializzati non ha portato ai miglioramenti sperati, non giustificando quindi il maggior sforzo richiesto.
Dalla realizzazione del presente caso d'uso le soluzioni SaaS e PaaS, ognuna nel proprio contesto di applicazione, risultano le scelte più indicate, andando quindi ad escludere la soluzione IaaS.

7 Conclusioni

In questo capitolo conclusivo è esposta un'analisi dei risultati del progetto trattato grazie alle fasi descritte nei capitoli precedenti, in modo da compararlo con i presupposti iniziali che hanno portato al presente lavoro.
Inoltre, sono presentati i possibili miglioramenti futuri del caso d'uso proposto al fine di migliorarne le funzionalità rese disponibili ed è riportata la continuazione che avrà il presente progetto nei prossimi mesi e anni all'interno dell'azienda in cui è stato realizzato.

In primo luogo, è possibile affermare che l'obiettivo di valutare l'impatto che servizi cloud di intelligenza artificiale abbiano nella diffusione e nella possibilità di applicazione di tali processi nelle organizzazioni sia stato portato a termine, mostrando come questi servizi, nelle declinazioni trattate, semplifichino ampiamente l'adozione di tali tecnologie, abbattendo le barriere d'ingresso, sia in termini di costi da sostenere che in termini di competente necessarie. Come presentato nel capitolo precedente, infatti, la realizzazione di una soluzione, declinata nelle tre metodologie di erogazione del cloud, non abbia richiesto eccessive competenze nell'ambito dell'intelligenza artificiale e che non sia correlata ad un eccessivo investimento iniziale, ma che, al contempo, permetta di sfruttare un innovativo strumento di AI in modo personalizzato ed efficiente. L'utilizzo di servizi dedicati a specifiche funzionalità, connessi tra loro, ha permesso di realizzare una soluzione standardizzata di uno strumento di chatbot, integrabile in moltissimi contesti. È infatti possibile, fornire questo tool particolari limitazioni sul caso d'uso specifico e sui dati forniti. Infine, come mostrato, risulta di fondamentale importanza la scalabilità del cloud, che ci ha permesso di realizzare strutture modulari, che si ripercuotono su costi contenuti per gli utenti finali.

L'analisi del mercato dei servizi cloud disponibili sulla piattaforma Azure AI, così come i requisiti emersi dal confronto con gli stakeholders, hanno fornito vari spunti per l'estensione delle funzionalità offerte. Infatti, è emerso l'interesse di

poter fornire allo strumento ulteriori fonti di dati, non esclusivamente testuale, in particolare la possibilità di convertire la base dati di video di formazione creata da molte organizzazioni, in dati accessibili dallo strumento analizzato; questa estensione ha trovato immediatamente applicabilità teorica tramite il servizio Video Analyzer, disponibile nella piattaforma, che permette di convertire un video nella propria trascrizione e descrizione. Ulteriori estensioni potrebbero riguardare compiti specifici gestibili dallo strumento, come per esempio la restituzione di informazioni necessarie all'utente da inserire in un form da compilare, compreso il significato di quest'ultimo e avendo conoscenza dell'utente dal quale giunge la richiesta, esso potrebbe fornire direttamente tali dati.

Il progetto condotto presso Cluster Reply, è stato qui trattato in maniera parziale, ovvero non sono stati riportati tutti i casi d'uso e le loro declinazioni realizzate. Il presente elaborato si pone come obiettivo fornire i concetti fondamentali per comprendere l'impatto di queste tecnologie nel contesto aziendale e di come il cloud possa essere il driver principale per la adozione, d'altronde il caso d'uso mostrato ci ha permesso di comprendere a fondo le potenzialità e applicazioni dell'AI integrata al cloud computing.
Il progetto non si è interrotto infatti, con il caso d'uso presentato: al momento della redazione di questo elaborato sono stati realizzati ulteriori casi d'uso, anche secondo diverse infrastrutture, in modo da comprendere sempre più a fondo le possibilità fornite dai cloud provider ai system integrator e alle organizzazioni.
I risultati di questo studio sono stati fondamentali nelle prime fasi del progetto, in particolare l'analisi delle caratteristiche delle tre principali tipologie di erogazione dei servizi cloud è risultato di assoluta rilevanza per la progettazione delle infrastrutture delle soluzioni, infatti considerata l'analisi comparativa precedentemente mostrata, è stato cruciale comprendere la correlazione tra grado di personalizzazione della soluzione, costi di realizzazione e qualità del risultato. Inoltre, l'analisi dei cloud service provider e in particolare della piattaforma Azure AI ha permesso di avere un quadro chiaro e ampio delle funzionalità disponibili e di come sfruttarle nei diversi contesti.

Bibliografia e sitografia

[1] K. Haan e R. Watts. "How Businesses Are Using Artificial Intelligence In 2023". In: *Forbes* (2023).

[2] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. *The AI Index 2023 Annual Report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 2023.

[3] Lins, S., Pandl, K.D., Teigeler, H. "Artificial Intelligence as a Service". In: (2021).

[4] IBM. *IBM Global AI Adoption Index*. IBM, 2022.

[5] *Cloud AI Market size and Share Analysis - Growth Trends forecasts (2023 - 2028)*. Mordor Intelligence, 2023.

[6] McKinsey & Company. *The state of AI in 2022 - and a half decade in review*. McKinsey & Company, 2022.

[7] KPMG. *2023 Generative AI Adoption Index*. KPMG, 2023.

[8] Gartner Research. *Gartner Magic Quadrant for Cloud Infrastructure and Platform Services*. Gartner, 2022.

[9] OpenAI. *OpenAI and Microsoft extend partnership*. URL: `https://openai.com/blog/openai-and-microsoft-extend-partnership`.

[10] Gartner Research. *Gartner Forecasts Worldwide Public Cloud End-User Spending to Reach Nearly $600 Billion in 2023*. Gartner, 2023.

[11] M. Shirer. *Worldwide Public Cloud Services Revenues Surpass $500 Billion in 2022, Growing 22.9% Year Over Year, According to IDC Tracker*. International Data Corporation (IDC), 2023.

[12] J. Roach. "How Microsoft's bet on Azure unlocked an AI revolution". In: (2023).

[13] *Gartner*. URL: `https://www.gartner.com/en/about`.

[14] *Amazon Web Services*. URL: https://aws.amazon.com/it/.

[15] *Google Cloud Platform*. URL: https://cloud.google.com/?hl=it.

[16] *Microsoft Azure*. URL: https://azure.microsoft.com/it-it.

[17] *OpenAI*. URL: https://openai.com/about.

[18] *Grammarly*. URL: https://www.grammarly.com/.

[19] *Midjourney*. URL: https://www.midjourney.com/.

[20] *GitHub Copilot*. URL: https://github.com/features/copilot.

[21] *AI Builder*. URL: https://powerapps.microsoft.com/en-us/ai-builder/.

[22] *Dynamics 365 Copilot*. URL: https://blogs.microsoft.com/blog/2023/03/06/introducing-microsoft-dynamics-365-copilot/.

[23] *Cluster Reply - About Us*. URL: https://www.reply.com/cluster-reply/it/about-us.

[24] *Reply*. URL: https://www.reply.com/it.

[25] *Azure Pricing Calculator*. URL: https://azure.microsoft.com/it-it/pricing/calculator/.