OVER DATA.

Artificial Intelligence
& Tech Culture

INDICE

01

Il caso Europcar

Fleet asset capacity e revenue management optimization attraverso advanced prescriptive analytics

Guidare un'azienda altamente dinamica, in un mercato competitivo è piuttosto impegnativo. Questa sfida è anche associata ad elevati rischi finanziari quando l'azienda richiede investimenti importanti. È il caso della multinazionale Europcar nel settore del noleggio auto. Questo articolo mostra come gli advanced analytics possono essere opportunamente combinati e adattati per migliorare il business management. La corretta integrazione di software math-based può aiutare le aziende a ridurre i costi e guadagnare quote di mercato in un contesto complesso e competitivo.

L'articolo illustra gli elementi di un Revenue Management innovativo che combina analisi predittive e prescrittive. La previsione della domanda è pensata come una spina dorsale che fornisce una visione unica e condivisa dei ricavi ai decisori logistici, le cui scelte sono supportate dall'ottimizzazione matematica. Dal punto di vista della

> **L'attività è influenzara da variabilità, casualità, stagionalità, eventi speciali, condizioni meteorologiche e dalla concorrenza**

gestione aziendale, il contributo illustra quali condizioni il top management dovrebbe creare per ottenere risultati di successo attraverso advanced analytics.

Il contesto di business del rental car

I servizi vengono forniti ai clienti da stazioni di noleggio distribuite in un determinato territorio, tipicamente un intero paese. Una stazione di noleggio auto locale è il luogo fisico in cui un cliente ritira un'auto e la riconsegna al termine del servizio. Le stazioni possono essere eventualmente raggruppate in zone, mentre le auto vengono divise per categoria in base al loro livello, ed i clienti vengono suddivisi in segmenti (leisure, business, ecc.).

La governance del business dell'autonoleggio richiede una combinazione di decisioni su più intervalli di tempo. L'obiettivo dei decisori è avere le auto giuste, nelle stazioni giuste, al momento giusto, e venderle al giusto prezzo.

Le decisioni che incidono sulle entrate possono essere raggruppate nelle seguenti categorie:

1. **ASSET MANAGEMENT:**

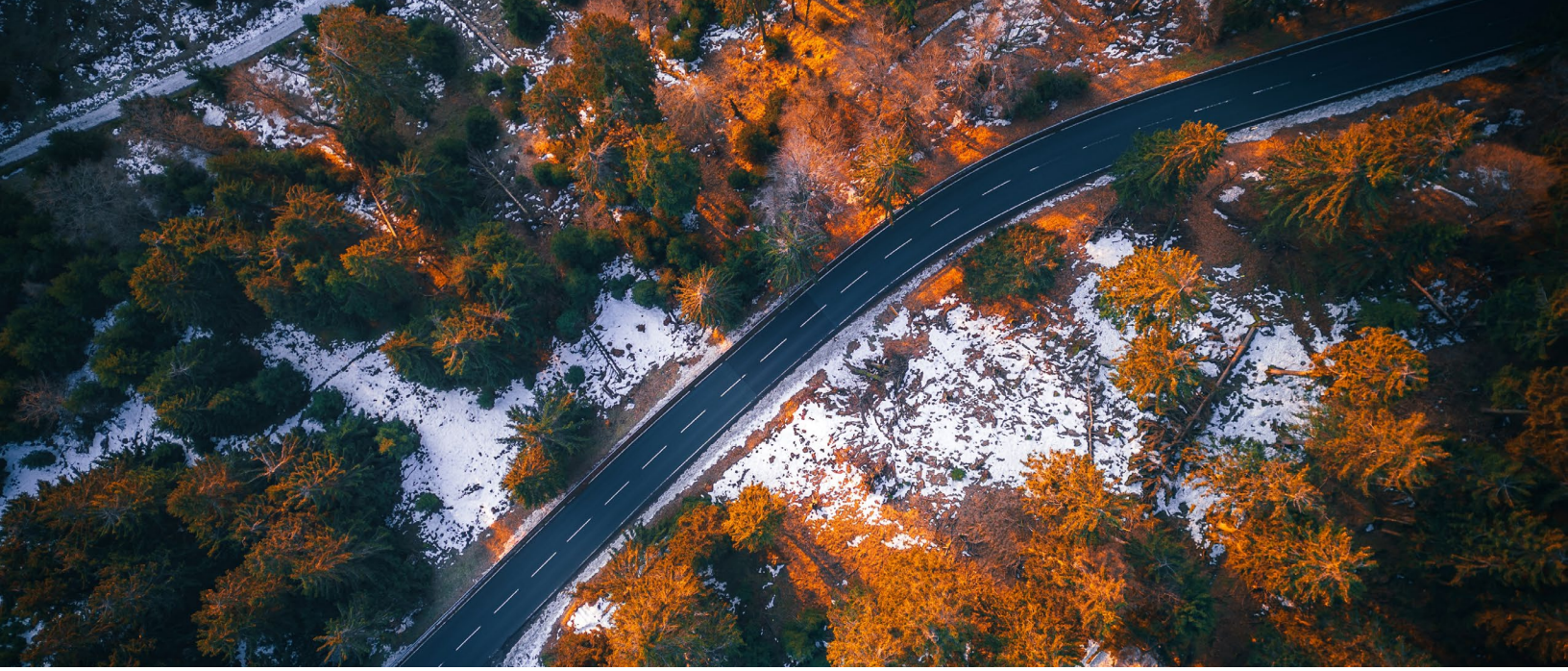
- **Lungo termine (12 mesi):** definire i contratti con i venditori di automobili stabilendo il numero di auto da acquisire e dismettere (restituire), nonché le condizioni economiche comprese le penali nel caso in cui le auto superino il chilometraggio massimo predefinito;
- **A medio termine** (pochi mesi): pianificare l'ingresso e lo scarico delle auto da/verso le stazioni giuste;

2. LOGISTIC MANAGEMENT – MEDIO O BREVE TERMINE:

Pianificare i trasferimenti delle auto tra le stazioni per coprire i picchi di domanda o limitare il basso utilizzo della flotta in determinate zone;

3. COMMERCIAL MANAGEMENT – MEDIO O BREVE TERMINE:

- **Definire il prezzo per ogni auto (gruppo) e stazione** in base alla durata del noleggio e alle condizioni di mercato previste;
- **Definire le promozioni;**
- **Bloccare le vendite** in caso di domanda eccessi-



va prevista.

Il problema

L'attività è influenzata da variabilità, casualità, stagionalità, eventi speciali come fiere, festività, condizioni meteorologiche insolite (come la neve) e, ultimo ma non meno importante, dalle azioni della concorrenza. Tutti questi fattori determinano i risultati aziendali finali (il profitto). Di conseguenza, qualsiasi decisione commerciale, riguardante sia la gestione della flotta che la gestione delle vendite, espone i decisori a un certo rischio di creare una bassa saturazione (deterioramento dei profitti) o una perdita di opportunità commerciali (l'indisponibilità di auto nella stazione corretta). Ad ogni azione di gestione sono associati determinati costi e risultati. Ad esempio, trasferire alcune auto da una stazione all'altra per intercettare la domanda dei clienti, comporta un costo che deve essere considerato nel bilancio complessivo dei ricavi.

Il capacity management è fortemente connesso con l'ambito commerciale e il marketing, e viceversa. Le decisioni in un dipartimento possono influenzare in modo significativo le decisioni di un altro. Di conseguenza, più decisori devono condividere una visione globale e sincronizzare delle proprie azioni per raggiungere un obiettivo comune: ottenere il massimo profitto dalle risorse disponibili. Questo è l'obiettivo del Revenue Management.

Il revenue management e le principali complessità per l'attività di noleggio auto

Revenue management nel settore delle auto a noleggio presenta somiglianze con quella delle compagnie aeree e del settore alberghiero. Il Revenue Management nel settore del rental car condivide, con i casi della compagnia aerea e dell'hotel, diversi elementi: domanda prevedibile, inventario deperibile, struttura adeguata di costi e prezzi e domanda variabile nel tempo. D'altra parte, una caratteristica significativa del contesto del Revenue Management del noleggio auto è la natura della capacity (automobili). La capacity è molto più flessibile rispetto al Revenue Management delle compagnie

aeree o degli hotel, poiché la dimensione della flotta può variare nel tempo. Per questi motivi, anche se per il Revenue Management delle compagnie aeree e alberghiere esiste un'ampia letteratura a disposizione, per il business del noleggio auto la letteratura è meno sviluppata.

La soluzione per Europcar

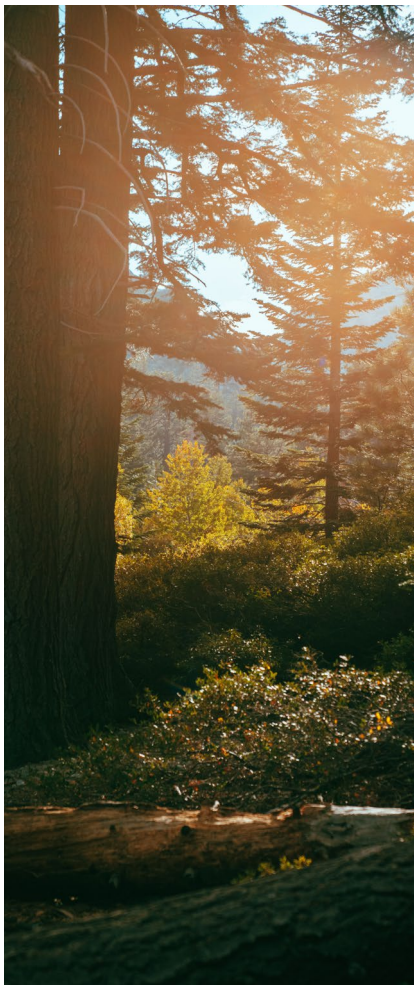
Il Gruppo Europcar è uno dei principali attori nei mercati della mobilità ed è quotato su Euronext Paris. La missione del Gruppo è quella di rappresentare un'alternativa interessante all'auto di proprietà fornendo un'ampia gamma di soluzioni di mobilità: noleggio auto, Vans & Trucks, servizio con conducente, car sharing o peer-to-peer.

Europcar è il marchio principale del Gruppo e offre il noleggio di veicoli a breve termine. Europcar offre un'ampia varietà di modelli recenti di autovetture e veicoli commerciali leggeri per il noleggio su base giornaliera, settimanale o mensile. Per estendere ed espandere la propria presenza a livello mondiale, Europcar si avvale anche di partnership e accordi generali di rappresentanza commerciale. Le agenzie di noleggio gestite direttamente da Europcar sono situate nei paesi in cui l'azienda vanta una presenza pluriennale e una vasta esperienza, ovvero Germania, Australia, Belgio, Danimarca, Spagna, Francia, Irlanda, Italia, Nuova Zelanda, Portogallo e Regno Unito

.

La storia di Europcar

Europcar viene fondata nel 1949 a Parigi da Raoul-Louis Mattei con il nome di "L'Abonnement Automobile". Il nome Europcars appare per la prima volta nel 1951. Nel 1974 viene eliminata la "s" finale e il nome dell'azienda diventa Europcar e vengono create nuove filiali in Spagna, Gran Bretagna, Italia e Portogallo. Europcar lancia un nuovo servizio di prenotazione online nel 2001. Questo è un grande passo avanti per Europcar. Il suo sito web www.europcar.com, così come i siti web locali, fornisce 200.000 veicoli che compongono la sua flotta in 118 paesi. I noleggi orari e i pagamenti anticipati Europcar, precedentemente riservati alle compagnie aeree, sono ora disponibili per tutti i clienti. Nel 2003 Europcar diventa leader europeo nel noleggio di veicoli grazie ad una strategia basata su numerosi affiliati e sullo sviluppo di numerose partnership con agenzie di viaggio, compagnie aeree, ecc. Cinque anni dopo, nel 2008, viene formata una nuova alleanza strategica con il leader nordamericano Enterprise Rent-a-Car, con l'obiettivo di continuare ad espandersi. Il gruppo conta più di 8.000 filiali e una flotta di oltre 1 milione di veicoli. Nel 2009, Europcar implementa un servizio di prenotazione mobile, che consente ai clienti di individuare la



filiale più vicina e prenotare un veicolo utilizzando il proprio telefono cellulare. Questo servizio innovativo, che garantisce maggiore mobilità ai clienti, è operativo su tutta la rete internazionale di Europcar. Nel 2012, il gruppo lancia un ampio programma di trasformazione, noto come "Fast Lane", per aumentare la presenza del gruppo sul mercato e aprire la strada alla sua transizione da società di noleggio veicoli a operatore di servizi di mobilità, e migliorare ulteriormente la propria operatività. La soddisfazione del cliente è fondamentale per il Gruppo e per i suoi dipendenti: questo impegno spinge al continuo sviluppo di nuovi servizi. Europcar Lab, ad esempio, è stato creato per comprendere meglio le sfide future della mobilità attraverso l'innovazione e investimenti strategici come Ubeeqo ed E-Car Club.

Per rispondere meglio alle nuove esigenze dei clienti nel campo della mobilità, il Gruppo continuerà a sfruttare le sue risorse chiave, vale a dire le sue capacità di gestione della flotta e la sua attività diversificata portata del cliente. L'azienda ritiene che sfruttare la propria rete come servizio a terzi sia un'importante opportunità per rafforzare la propria dimensione. Inoltre, l'accelerazione della trasformazione digitale attraverso la modernizzazione della propria rete consentirà al Gruppo di migliorare il percorso dei propri clienti e di attingere sempre più ai segmenti in più rapida crescita dell'ecosistema della mobilità.

Come Europcar ha affrontato l'evoluzione del mercato

Prima di Opticar, Europcar non disponeva di uno strumento standardizzato per quanto riguarda l'attività di Revenue and Capacità Management (RCM). I nove paesi in cui operava Europcar avevano i propri strumenti e processi locali, che utilizzavano

> La previsione diventa la soluzione consentendo al decisore di condivere una visione futura comune

per seguire l'attività e cercare di ottimizzare ricavi e capacità. Da un punto di vista locale, ha comportato una grande quantità di lavoro manuale, aggregando dati eterogenei provenienti da più strumenti operativi, in molti formati diversi. Nella maggior parte dei casi, questi dati venivano mescolati in enormi file Excel utilizzando macro e calcoli di grandi dimensioni. A causa del corrispondente carico di lavoro sostenuto, alla fine il risultato è stato l'utilizzo di diversi file Excel, aggiornati al massimo due volte a settimana, per prevedere e monitorare la domanda e la disponibilità della flotta. La domanda è stata gestita e prevista solo come volume complessivo. I team di revenue e capacity management hanno esaminato il Business On the Books (BOB) su una base a brevissimo termine e principalmente da un punto di vista dell'affitto, non da un punto di vista del Revenue Management. La previsione manuale poteva essere limitata solo a pochi oggetti. L'attenzione era posta sulle attività svolte nell'ultima settimana e confrontate con le tendenze costruite in base alla conoscenza degli utenti, alla stagionalità, alle prestazioni attuali, ai dati storici e alle stime sulle nuove attività. Infine, gli analisti dei prezzi controllavano manualmente il posizionamento dei prezzi sui siti web dei concorrenti e dei broker. Non veniva utilizzato alcuno strumento standard per avere in un'unica visualizzazione le informazioni su domanda, capacità e prezzi allo stesso tempo.

Per gli analisti di Opticar, anche consolidare i dati e guidare l'attività a livello di Gruppo con un tale livello di eterogeneità è stata una sfida. In un mercato così competitivo come quello dell'autonoleggio, il revenue e capacity management è stato posto in modo del tutto naturale al centro dei processi di Europcar. Avere visibilità della domanda (contratti di noleggio aperti, prenotazione, domanda prevista, ecc.) per gruppi di veicoli diventa un must, per consentire ai team RCM di essere più dettagliati nella loro strategia e sfruttare opportunità di guadagno e profitto attraverso gruppi di veicoli e canali di distribuzione. Senza questo livello di dettaglio, era quasi impossibile identificare i segmenti redditizi e implementare strategie di rendimento accurate. Avere una previsione della domanda prevista, calcolata per regione, zona, gruppo di veicoli per le 24 settimane a venire si è rivelato obbligatorio per dare una visione accurata del tasso di utilizzo previsto. Pertanto, nasce la necessità di prendere decisioni strategiche sui prezzi, con sufficiente anticipo per

beneficiare della situazione del mercato. Non ultimo, è diventata evidente anche la necessità di disporre di uno strumento standard, dove le informazioni fossero disponibili online in qualsiasi momento da tutti gli utenti, a livello locale e centrale, condividendo la stessa lingua e gestendo le stesse informazioni. Opticar è stata quindi progettata per rispondere a tutte queste esigenze.

Opticar, la soluzione di revenue management per Europcar

Opticar supporta decisioni molteplici e interrelate basate sul demand forecast. La previsione diventa la spina dorsale della soluzione consentendo al decisore di condividere una visione futura comune. Con il termine domanda si fa riferimento al numero di ritiri di auto in un dato intervallo di tempo, legato ad una combinazione di stazione (o zona), a un segmento di clientela e a una tipologia di veicolo.
La previsione è richiesta per più di venti settimane e si differenzia in base ai seguenti driver:

1. Geografia (stazioni e zone);
2. Tipologia di vetture;
3. Segmento di clientela;
4. Tempo (dati storici).

Nel breve termine, infatti, è necessaria una previsione giornaliera mentre, nel medio termine, una previsione settimanale può essere opportuna per supportare la maggior parte delle decisioni.
Sono inoltre da evidenziare i seguenti fenomeni che il sistema deve modellare:

- Cancellazione delle prenotazioni esistenti;
- Up sell al desk (il cliente, ad esempio, ha prenotato una Fiat Panda, ma il venditore al desk
- vende un upgrade alla Mini);
- Possibilità di prolungare un contratto esistente/ aperto;
- ''One Way'' (la stazione del check-in è diversa dalla stazione del check-out).

Il software Opticar produce automaticamente due diversi tipi di previsioni, calcolate ogni giorno:

1. La previsione non vincolata: si presuppone una disponibilità illimitata di risorse (automobili);
2. La previsione vincolata: vengono prese in considerazione le limitazioni di capacità.

I modelli previsionali sono in grado di tenere conto di fattori d'influenza come festività, eventi speciali (ad es. fiere) e condizioni meteorologiche. Successivamente un sofisticato modello di simulazione elabora la previsione per calcolare i ricavi e l'utilizzo della flotta nel tempo. Gli indicatori di performance dipendono da altri fattori che il sistema deve considerare, come la durata del noleggio o gli accessori inclusi (GPS, assicurazioni, ecc.). In particolare, al fine di integrare l'analisi predittiva e quella prescrittiva, ogni notte la previsione viene utilizzata per creare un'analisi vincolata e non vincolata (''scenari di base''). Gli scenari di base vengono utilizzati come input per i moduli di ottimizzazione matematica (analisi prescrittiva). L'ottimizzatore della flotta, considerando tutte le componenti di costo, identifica la soluzione ottimale in termini di trasferimenti di auto tra stazioni, auto in

flotta o in scarico, che minimizza il costo complessivo e massimizza le entrate adattando al meglio la domanda. Utilizzando questi scenari di base come punto di partenza, i decisori possono anche valutare, tramite analisi what-if, l'impatto di eventuali azioni e confrontare gli scenari. Ad esempio, è possibile confrontare diverse alternative in termini di trasferimenti di auto tra le stazioni. Un'altra variabile importante è il prezzo. I prezzi devono essere decisi in base alla domanda, alla posizione di mercato e alle risposte della concorrenza. La scelta del livello di prezzo ottimale è un'altra caratteristica importante nella gestione dei ricavi delle auto a noleggio.
L'importanza della gestione del cambiamento.

Tecnologie e competenze tecniche, tra cui efficienza computazionale e architettura hardware, sono sicuramente ingredienti importanti per l'analisi predittiva e prescrittiva applicata ma, non meno importanti, sono i processi di gestione del cambiamento e l'impegno del top management per il successo di un progetto come questo. L'esperienza Europcar evidenzia chiaramente l'importanza di tale aspetto. Il top management di Europcar ha seguito tutte le fasi del progetto, compresa la reingegnerizzazione del processo, spingendo per un approccio decisionale collaborativo basato su dati e metodi oggettivi.
Inoltre, tempo qualificato è stato dedicato alla progettazione di corsi di formazione interni e alla guida dei manager operativi in tutti gli uffici nazionali in modo che loro e i loro team possano adattarsi a questo nuovo approccio. Parte integrante del programma di formazione continua messo in atto è anche la costante condivisione di esperienze e best practice tra i revenue manager.
Lo sforzo è stato ripagato in

> **Opticar ha migliorato l'attività quotidiano di Europcar nel processo decisionale rendendola più reattiva**

termini di business, visto che Europcar vede un incremento dei volumi, soprattutto in bassa stagione, ed anche in alta stagione, oltre ad un miglioramento delle tariffe medie (Revenue per Day). Inoltre, si risparmia tempo man mano che i report manuali cessano di esistere e i team possono concentrarsi sulle attività che aumentano le entrate, piuttosto che sul reporting manuale. La distribuzione delle automobili è migliorata, il che porta non solo a un migliore utilizzo della flotta, ma anche a una riduzione dei costi. Nel complesso, lo sviluppo e l'implementazione di OptiCar hanno avuto molto successo e lasciano ancora molte opportunità di guadagno da esplorare in futuro.

Come Opticar ha migliorato le operazioni quotidiane

Con Opticar, i team RCM di Europcar sono passati da un processo manuale a una soluzione automatizzata, affidabile e completamente integrata per supportare le decisioni aziendali e sui prezzi. Opticar ha rappresentato sicuramente un grande passo avanti per Europcar, non solo consentendo di gestire al meglio le sfide poste dalla sempre crescente complessità degli scenari di mercato, ma anche (e soprattutto) valorizzando la cultura della domanda rispetto alle dinamiche di capacità tra tutti i dipartimenti dell'azienda.

Opticar ha migliorato l'attività quotidiana di Europcar nel processo decisionale rendendola più reattiva rispetto al passato e standardizzando il modo in cui le informazioni vengono condivise. Ciò non solo all'interno delle funzioni RCM, ma anche con tutti gli altri stakeholder (operazioni, marketing e vendite) a livello di Gruppo. Opticar ha cambiato il modo in cui gli utenti Europcar guardano alle operazioni, offrendo in ogni momento una visione in uno strumento unico della disponibilità della flotta, degli affari in portafoglio, della concorrenza di mercato e, cosa più importante, della domanda prevista. La sua introduzione ha permesso di controllare quotidianamente l'andamento delle prenotazioni e degli affitti. Con il livello di dettaglio offerto dallo strumento, i gestori delle entrate e della capacità possono prendere le decisioni pertinenti al fine di ottimizzare l'allocazione della flotta tra le regioni, considerando la domanda prevista e garantire la capacità della flotta per i segmenti più redditizi. Poiché Opticar consolida molti report precedenti, riduce significativamente il tempo utilizzato per creare report ad hoc, nonché il carico di lavoro precedentemente impiegato nel controllo di tali report. Infine, i team RCM sono in grado di pianificare in anticipo la propria strategia e reagire rapidamente se è necessario un aggiustamento.

Dal punto di vista del Gruppo, Opticar ha consentito di allineare i processi di nove paesi aziendali su standard condivisi consentendo un'analisi approfondita e continua e supportando il business al fine di guidare in modo efficiente l'attività a livello di Gruppo.
Opticar come strumento aziendale nasce nel 2014 dall'adattamento di uno strumento pilota italiano in una soluzione RCM a livello di gruppo. Oltre a questi costi di implementazione, Europcar ha investito in un significativo piano di gestione del cambiamento che è stato implementato e seguito parallelamente in nove paesi aziendali, al fine di accompagnare, aiutare e supportare i team RCM nel loro processo di passaggio all'utilizzo di Opticar
Dal 2015 Europcar ha investito in media 250mila euro all'anno negli sviluppi IT al fine di potenziare la piattaforma Opticar, tenendo in considerazione il feedback degli utenti per migliorare l'usabilità di Opticar, oltre a migliorare costantemente la capacità del sistema di prendere in considerazione nuovi parametri e migliorare l'accuratezza delle previsioni.
Per garantire che tutte le parti interessate e gli utenti di Opticar condividano lo stesso livello di informazioni e conoscenze sullo strumento, ogni nuova versione include un aggiornamento del corpus della documentazione (specifiche tecniche, guide utente, casi d'uso ed esempi) e formazione online dedicata. Vengono organizzate sessioni con tutti i rappresentanti dei paesi.

Conclusioni

La metodologia Opticar è un'integrazione tra tecniche di previsione, simulazione e ottimizzazione. Questa integrazione sequenziale crea uno degli strumenti più avanzati per il settore dell'autonoleggio. Opticar ha cambiato il modo in cui gli utenti affrontano le operazioni quotidiane in Europcar. Fornendo un punto di partenza per gli utenti, è diventato il collegamento tra culture e metodologie di lavoro diverse. Ciò migliora la qualità delle informazioni con cui gli utenti possono iniziare a lavorare e il focus delle operazioni quotidiane è stato spostato sulle esigenze dei clienti. Gli azionisti hanno tratto la maggior parte dei benefici da Opticar. Molti sono gli impatti in tutti e nove i paesi aziendali che hanno migliorato le performance e i ricavi negli ultimi anni. Altri stakeholder, come i tour operator, hanno sottoscritto contratti con Europcar, per la sua affidabilità e la fiducia che i clienti ricevono da essa.
Infine, i clienti Europcar beneficiano di Opticar. Grazie ad un processo ottimizzato di gestione della capacità, Europcar è in grado di servire meglio i propri clienti, rispondendo alle loro esigenze di mobilità dove e quando conta di più. Analizzando più nel dettaglio la domanda composizione, il mix di auto può essere adattato alle aspettative dei clienti. In questo senso Europcar è in grado di fornire ad ogni cliente il veicolo giusto, a seconda delle sue necessità. Infatti, anche se ricevere un upgrade gratuito al desk può essere visto come un vantaggio, in alcuni casi, come quando si utilizza un veicolo in città, questo può essere uno svantaggio. Opticar ha inoltre contribuito a far sì che Europcar venisse più volte premiata come azienda leader nel noleggio auto.

I micro-cloud affrontano con successo le sfide dell’edge-computing

> **Nel “edge computing” l’elaborazione dei dati avviene sul “bordo” della rete**

La storia dell’evoluzione tecnologica aziendale è caratterizzata da una serie di tendenze e buzzword che spesso perdono rilevanza con il passare del tempo. Questo avviene sia perché alcune di queste tendenze sono state eccessivamente pubblicizzate, sia perché sono state soppiantate dalle nuove mode emergenti.

Attualmente, fra i concetti che fanno tendenza c’è quello di edge computing. L’ “Edge computing” è un paradigma di calcolo distribuito che comporta l’elaborazione dei dati e l’esecuzione delle computazioni in prossimità della fonte di generazione dei dati, anziché fare affidamento esclusivamente su server centralizzati in cloud o data center. Nell’“edge computing”, l’elaborazione dei dati avviene sul “bordo” della rete, tipicamente presso o vicino ai dispositivi o ai sensori che raccolgono i dati.

La crescente importanza dell’edge computing e la necessità per le aziende di sviluppare una solida strategia per sfruttare appieno questo modello si spiegano tenendo presente la possibile evoluzione dello scenario relativo all’ecosistema dei dati. Gartner prevede che, entro il 2025 si determinerà un cambiamento sostanziale nell’elaborazione dei dati aziendali. Secondo le stime dell’autorevole società americana, entro quella data ben il 75% dei dati generati verrà processato al di fuori dei data center tradizionali, siano essi on premise o concepiti per l’erogazione di servizi cloud.

Questa decentralizzazione dell’elaborazione dei dati ha il potenziale per trasformare praticamente tutti i settori, dall’industria delle telecomunicazioni a quella della produzione, influenzando persino l’esperienza di acquisto dei consumatori nei negozi fisici.

Le ragioni economiche alla base dell’adozione dell’edge computing sono molte. In primo luogo, riduce i costi legati alla trasmissione di dati a un cloud centrale, contribuendo a ottimizzare la larghezza di banda e riducendo le spese di archiviazione nei data center centrali. Inoltre, offre alle aziende un maggiore controllo sulla gestione dei dati, migliorando la resilienza organizzativa e consentendo una governance più precisa e conforme alle normative vigenti. Infine, l’edge computing apre la porta a nuovi servizi ad alto valore, grazie alla possibilità di offrire esperienze utente altamente personalizzate e sfruttare al massimo i dati locali, come quelli provenienti dai sensori.

LE SFIDE DELL'EDGE COMPUTING: UN NUOVO PARADIGMA PER LE IMPRESE

Tuttavia, l'adozione di una strategia di edge computing su larga scala comporta una serie di sfide cruciali:

1. **La necessità di adattarsi a esigenze in continua evoluzione.** Le applicazioni e le esigenze dei consumatori cambiano nel tempo, e il sistema di edge computing deve essere flessibile e adattabile. La scalabilità è un aspetto particolare da considerare. Mentre è possibile iniziare con pochi siti, magari un singolo sito di prova con quattro nodi (server), la vera sfida sorge quando è necessario replicare e scalare fino a centinaia o migliaia di siti. In questo scenario, una soluzione altamente scalabile è fondamentale.

2. **L'accesso fisico ai siti edge.** Dato che questi luoghi sono spesso distribuiti in località remote e difficili da raggiungere, è essenziale ridurre al minimo la necessità di interventi umani sul campo. Ciò richiede una soluzione altamente automatizzata con manutenzione in loco ridotta o addirittura nulla. Inoltre, le implementazioni di edge computing non sostituiranno necessariamente gli investimenti passati in cloud e data center privati. Pertanto, è fondamentale che un'architettura di edge computing sia in grado di interfacciarsi con l'infrastruttura esistente, che può comprendere sia cloud pubblici o privati sia

infrastrutture bare metal. Dato che molte delle infrastrutture, delle pratiche e delle politiche esistenti sono centralizzate, le soluzioni di calcolo decentralizzato con una gestione centralizzata sono più facili da integrare. Inoltre, sono necessarie API aperte e standard per garantire la piena interoperabilità con l'infrastruttura esistente.

3. **Essere a prova di futuro.** In un mondo in cui software e hardware sono in costante evoluzione, è essenziale evitare di diventare dipendenti da una tecnologia o da un fornitore specifico. Un buon livello di astrazione può contribuire a ridurre i rischi legati alla riconfigurazione o alla migrazione della tecnologia. Ad esempio, i micro-cloud devono essere indipendenti dall'hardware e fornire API standard per proteggere le applicazioni da modifiche esterne.

In sintesi, l'adozione di una variante localizzata delle tecnologie dei data center o dei servizi cloud rappresenta una potente strategia per superare queste sfide. Questo approccio consente alle aziende di sfruttare le tecnologie e le competenze che hanno già acquisito internamente, fornendo al contempo una piattaforma modulare e standard per la creazione di un ecosistema di siti decentralizzati.

LA PROMESSA DEI MICRO CLOUD: POTENZIARE L'EDGE CON INTELLIGENZA

Tuttavia, nonostante i vantaggi e il potenziale dell'edge computing, il termine stesso è spesso vago e può essere interpretato in modi diversi da diverse persone. Questo lo rende meno utile per le aziende che devono gestire numerosi "punti di estremità" all'interno della propria organizzazione. Ad esempio, il modo in cui un rivenditore utilizza sensori per automatizzare la gestione della sua flotta di distribuzione in tutto il mondo è molto diverso dal modo in cui una società di telecomunicazioni sfrutta l'edge computing per fornire servizi di rete 5G.

Per i leader aziendali sarebbe più utile un approccio che superi la mera etichetta dell'edge computing e descriva ciò che accade all'estremità della rete in un contesto più ampio, cioè un continuum di elaborazione aziendale. Questo continuum può comprendere il cloud pubblico, il cloud privato, una nuova categoria definita come micro-cloud e l'IoT. In effetti, gran parte di ciò che è noto come edge computing può essere raggruppato sotto le categorie di micro-cloud e IoT.

I "micro cloud" emergono come una soluzione di notevole rilevanza poiché combinano l'elasticità tipica dei cloud pubblici e privati con la sicurezza, la privacy, la governance e la bassa latenza necessarie negli ambienti di edge computing decentralizzati. I micro-cloud replicano le API e le funzionalità dei principali cloud su scala "edge", creando un ecosistema di sistemi interconnessi su una rete complessa.

I "micro cloud" possono essere descritti come una nuova classe di infrastrutture per l'elaborazione su richiesta all'edge. Si distinguono dall'Internet delle cose (IoT) in quanto non solo raccolgono dati, ma eseguono anche l'elaborazione di questi dati. In pratica, un micro-cloud è un cluster di nodi di calcolo con archi-

> La presenza di API aperte e standard è un altro vantaggio dei micro-cloud.

viazione e rete locali. Questo stack è composto da tecnologie resilienti, autorigeneranti e opinionated, che consentono di creare implementazioni ripetibili. Quando parliamo di tecnologie "opinonated", in particolare, ci riferiamo a un approccio che tende a limitare gli spazi di manovra dello sviluppatore e ad orientare pesantemente le sue scelte. Inoltre, i micro-cloud possono essere scalati verticalmente per adattarsi alle esigenze aziendali, il che li rende altamente versatili. Possono essere collocati in una vasta gamma di località, da supermercati a negozi, da stadi a filiali aziendali o persino alla base di antenne 5G. Questa vicinanza ai consumatori è un elemento chiave, poiché consente un'esperienza più fluida e reattiva.
Con il loro stack di componenti modulari è possibile personalizzare il ridimensionamento a ogni livello, dai server fisici alle macchine virtuali e ai container. In questo modo, le aziende possono adattare le proprie risorse informatiche alle mutevoli esigenze. Inoltre, i micro-cloud sono progettati per l'automazione completa, riducendo al minimo le operazioni in loco e semplificando l'installazione e l'aggiornamento automatizzati. Questo è particolarmente importante date le sfide legate all'accesso fisico limitato alle località remote in cui spesso sono collocati. La presenza di API aperte e standard è un altro vantaggio significativo dei micro-cloud. Questa caratteristica permette di adattarsi facilmente alle diverse dimensioni e soluzioni all'edge. Date le differenze notevoli tra le infrastrutture delle aziende, questa flessibilità è fondamentale per garantire un'implementazione efficace.

Infine, il livello di astrazione è fondamentale per l'efficacia dei micro-cloud. Separando le applicazioni dall'hardware sottostante, questo approccio consente alle aziende di essere pronte per il futuro. Le tecnologie edge continueranno a evolversi, ma con una buona architettura di micro-cloud, l'aggiornamento o la sostituzione di parti dell'infrastruttura sarà meno dispendioso e problematico.

COME SPINDOX UTILIZZA I MICRO CLOUD

La genesi, il caso IaaS Spindox

Spindox, un pioniere nell'adozione di tecnologie edge e un noto innovatore, ha stretto una partnership strategica con Canonical, l'azienda dietro Ubuntu, il sistema operativo open source leader nel settore del cloud. Questa

collaborazione ha portato alla creazione del proprio micro-cloud marchiato Spindox, progettato per soddisfare le esigenze interne dell'azienda. Attraverso questo esercizio di dogfooding, Spindox ha migrato con successo più di un centinaio di macchine virtuali Windows e Linux su una nuova infrastruttura moderna e ad alte prestazioni. Questo ha non solo migliorato gli SLA e le performance complessive, ma ha anche contribuito a ridurre il consumo energetico e le relative emissioni di CO2. Il nuovo micro-cloud di Spindox è stato progettato, dimensionato e implementato in collaborazione con Canonical all'interno del progetto Exodus. Questa iniziativa non solo ha portato a un notevole successo nella migrazione e nell'aggiornamento tecnologico, ma ha anche consentito a Spindox di accreditarsi come partner ufficiale e di creare una practice specialistica a disposizione di clienti e prospect.

La naturale evoluzione di Spindox, il caso PaaS per OVS

L'esperienza acquisita nel contesto del progetto Exodus e la conseguente creazione della nuova Capability Micro-Cloud hanno rappresentato l'abilitatore della naturale evoluzione dell'offerta di Spindox per OVS, un cliente già utilizzatore della piattaforma Ublique di proprietà di Spindox.
Grazie alla sua architettura cloud-native, Ublique sfrutta appieno il micro-cloud, beneficiando delle avanzate funzionalità del cluster, tra cui autoscaling e autohealing, oltre alle più avanzate tecniche di disaster recovery e business continuity. Questo cluster è esteso a molti punti vendita distribuiti in tutto il territorio nazionale, garantendo il massimo livello di servizio.
Attualmente, stiamo conducendo analisi per estendere il perimetro funzionale del micro-cloud all'interno di OVS, oltre i confini dell'applicativo Ublique, seguendo l'esempio di successo di Spindox nell'integrare l'utilizzo del cloud all'interno della propria soluzione.

Il futuro dell'edge è micro

In conclusione, sebbene quello di edge computing sia un concetto importante e promettente, la sua presentazione sotto forma di continuum di elaborazione aziendale può riflettere in modo più accurato le esigenze delle aziende e semplificare la gestione di questa complessa realtà. Micro-cloud e IoT sono elementi fondamentali di questa equazione e devono essere considerati separatamente, ciascuno con le proprie sfide e opportunità.
Oggi i micro-cloud rappresentano l'apice dei progressi nel campo del cloud computing. Costituiscono un'innovazione che combina la flessibilità dei servizi cloud con la sicurezza e l'efficienza degli ambienti di edge computing decentralizzati. Questa tecnologia è destinata a beneficiare una vasta gamma di settori e applicazioni, ottimizzando le infrastrutture esistenti e consentendo nuove esperienze per i clienti finali. Tuttavia, è fondamentale che questa transizione sia supportata da una solida strategia a prova di futuro, data la costante evoluzione dell'edge computing.

03

Il PNRR investe sull'intelligenza artificiale: parte il progetto DESIGN-IT

Spindox annuncia l'ammissione formale del progetto DESIGN-IT da parte del Ministero delle Imprese e del Made in Italy ai finanziamenti previsti dal PNRR nell'ambito della Missione 4 (Istruzione e ricerca), Componente 2 (Dalla ricerca all'impresa). Il progetto (protocollo n. 69, Avviso Pubblico Accordi per l'Innovazione DM 31/12/2021 – I sportello), vede Spindox nel ruolo di mandataria e coinvolge altri tre attori: ISMN-CNR, Mister Smart Innovation e Reepack. Il totale dell'investimento dei quattro partner è pari a 5.392.557,50 euro, cui corrisponde un'agevolazione totale di 2.683.607,62 euro, di cui 1.619.859,38 euro a fondo perduto per Spindox. Tale contributo servirà a finanziare lo sviluppo di una piattaforma di supporto alle decisioni con largo impiego di tecniche di intelligenza artificiale.

> La piattaforma consentirà di progettare applicazioni basate sull'IA con un approccio zero-code

DESIGN-IT: Digital Twins a supporto delle decisioni

Il progetto DESIGN-IT (DEcision Science desIGN platform for digItal Twins) ha come obiettivo principale lo sviluppo di una piattaforma di supporto alle decisioni che utilizzi molteplici tecniche di intelligenza artificiale e applichi il concetto di Gemello Digitale (Digital Twin) in modo pervasivo, per aumentare le capacità degli operatori di prendere decisioni complesse.

Al contempo la piattaforma consentirà agli utenti di progettare e generare nuove applicazioni basate sull'intelligenza artificiale con un approccio zero-code, ovvero senza essere esperti di sviluppo software.
Un Digital Twin è un modello digitale in grado di recepire le caratteristiche fondamentali di un prodotto o di un processo, che ne replica fedelmente il funzionamento; è alimentato da un flusso di dati in tempo reale provenienti dalla controparte fisica generando un'effettiva copia digitale, sincronizzata con il prodotto o il processo di interesse, sulla quale poter effettuare simulazioni, ottimizzazioni, validazioni, pima di riportarle sul sistema reale.

In particolare, la piattaforma, che sarà validata nell'ambito della produzione di macchine e centri di lavoro per il manifatturiero, permetterà di:

1. Costruire Gemelli Digitali a più livelli, ovvero a livello sia di prodotto/macchinario o componente che di processo;

2. Mettere a disposizione dei Gemelli Digitali, e dunque delle controparti reali, servizi di Intelligenza Artificiale che ne ottimizzino il design e/o l'operazione;

3. Abilitare utenti non esperti di software alla costruzione di nuovi Gemelli Digitali e/o applicativi basati su Intelligenza Artificiale;

4. Aumentare l'adozione e incrementare la collaborazione uomo-macchina, grazie all'utilizzo di tecniche di Intelligenza Artificiale spiegabili, affidabili e robuste.

Il sistema verrà sviluppato secondo l'approccio dell'Agile Software Development, che a partire dagli anni Duemila ha preso piede come metodo più efficace ed efficiente per lo sviluppo software, permettendo di avere continui e rapidi rilasci "prototipali" di software per rendere partecipi tutti gli stakeholders dello stato di avanzamento del sistema in via di sviluppo.
Il progetto DESIGN-IT, iniziato a gennaio 2023, è attualmente in corso e si prevede di terminarlo entro dicembre

2025. Il gruppo di ricerca Spindox principalmente coinvolto elle attività è quello che fa capo ad aHead Research, il team che sviluppa e realizza soluzioni basate su Intelligenza Artificiale, sia per progetti interni che per partner e clienti.
Le competenze in campo sono molteplici e tutte di elevato livello, con risorse che si confrontano giornalmente con lo stato dell'arte dei modelli, dei metodi e delle tecnologie per poter concepire soluzioni che possano essere maggiormente efficaci ed efficienti e si possano qualificare come innovative sul mercato.

La piattaforma Design-IT si configurerà come un nuovo prodotto per il mercato, orientato principalmente alle aziende per le quali c'è difficoltà nel dotarsi di una serie di piattaforme o software differenti che richiedono di solito forti competenza in campo di programmazione e sono orientati ad esperti di simulazione e modellazione più che ad esperti di settore.
Gli strumenti analitici basati su Intelligenza Artificiale che saranno integrati nel nuovo prodotto, rispetteranno i criteri oggi richiesti in ambito di Trustworthy and Explainable AI, cioè algoritmi matematici i cui risultati possano essere spiegati ed interpretati e non semplicemente generati da un oracolo.
Una volta sviluppata la piattaforma, alla base del modello di business si prevede l'erogazione via SaaS (Sofware-as-a-Service) dei servizi che permetteranno di offrire soluzioni modulari e flessibili a un costo decisamente competitivo. Tra i servizi che potranno essere offerti sul mercato possiamo annoverare:

- manutenzione predittiva, questa estensione permette di sfruttare il Gemello Digitale del macchinario e

il flusso di dati raccolti in near-real-time per anticipare la necessità di operazioni di manutenzione e minimizzare il rischio di danneggiamenti alla macchina;

- previsione della domanda, il sistema che si svilupperà, basato su sistemi di previsione integrati a modelli di simulazione e ottimizzazione del Digital Twin e permetterà di superare gli attuali limiti dell'approccio predittivo basato solo sui dati storici.
- ottimizzazione per Eco-design, il sistema integrerà un ottimizzatore in grado di comunicare iterativamente con il Digital Twin del prodotto e macchinario per verificare l'impatto che modifiche al design della macchina o per prodotto per incrementare la potenzialità di circolarità (cioè la misura della possibilità di essere inserito all'interno di una catena del valore circolare). Algoritmi per gestione ottimale processi e catene logistiche sostenibili: il sistema includerà una serie di algoritmi che, sfruttando le informazioni raccolte dal Gemello Digitale del processo o della catena logistica è in grado di fornire supporto decisionale per massimizzare la sostenibilità del processo e la sua efficienza dal punto di vista energetico.

Un partenariato di spessore per un progetto innovativo

L'iniziativa vede Spindox nel ruolo di società mandataria. Spindox lavorerà in partnership con la sede di Bologna dell'Istituto per lo Studio dei Materiali Nanostrutturati del CNR (Consiglio Nazionale delle Ricerche), il consorzio MISTER Smart Innovation e la società Reepack.
L'Istituto per lo Studio dei Materiali Nanostrutturati (ISMN) è parte del CNR e afferisce al Dipartimento Scienze Chimiche e Tecnologie dei Materiali. Si tratta di una realtà riconosciuta a livello internazionale per le sue attività multidisciplinari nel campo dei materiali nanostrutturati e delle tecnologie e processi abilitanti. ISMN integra tecnologie abilitanti e con carattere trasversale che includono i materiali avanzati, la fotonica, la nanotecnologia, la biotecnologia e i processi chimici e manifatturieri avanzati.
MISTER Smart Innovation è una società consortile fondata nel 2009 e partecipate da istituzioni pubbliche quali CNR, Università degli Studi di Ferrara e Università di Parma, oltre a importanti aziende del territorio. Con sede nell'Area della Ricerca CNR di Bologna, MISTER è accreditato alla Rete Alta Tecnologia dell'Emilia-Romagna sia come Laboratorio di Ricerca Industriale, sia come Centro per l'Innovazione. Inoltre, MISTER è membro di quattro CLUST E-R regionali: Mech, Health, Innovate e Create.
Reepack è una PMI fondata nel 1997 con sede a Seriate (BG), specializzata nella vendita, progettazione e fabbricazione di macchinari per il confezionamento e soluzioni per l'automazione, applicate principalmente al settore alimentare. La vendita è orientata a un mercato internazionale con un export che incide per il 90-95% del fatturato.

Intelligenza artificiale

Un Viaggio nella Rivoluzione Tecnologica

2015

Nasce OpenAI

Elon Musk, Sam Altman, Peter Thiel e altri creano un laboratorio che ha come missione lo sviluppo si un'intelligenza artificiale generale. Lanciano **OpenAI Gym**, un sistema per addestrare l'IA attraverso ricompense, e successivamente **OpenAI Universe**, una piattaforma per misurare e addestrare l'IA su vari compiti, simili a quelli umani.

Pappagalli della probabilità

Secondo uno studio dell'Association for Computer Machinery (ACM), i modelli linguistici operano come **pappagalli della probabilità**, creando sequenze linguistiche basate sulla probabilità. Tuttavia, il testo sembra confuso e menziona la grandezza dei modelli linguistici senza chiarezza.

DEEPFAKE

Nel dicembre 2017, il termine "deepfake" è emerso quando un utente Reddit chiamato "Deepfake" ha iniziato a pubblicare video falsi, spesso a sfondo sessuale, utilizzando il deep learning per sostituire i volti di celebrità in contenuti esistenti. Questa tecnologia ha sollevato preoccupazioni sull'uso di video manipolati per diffondere disinformazione attraverso i social network.

2019

Microsoft investe in OpenAi

Elon Musk abbandona il progetto e OpenAI riceve 1 miliardo di dollari di investimento da **Microsoft**. Nasce un controversia con **ChatGPT 2** che genera notizie false. OpenAI allora chiede il monitoraggio governativo per tecnologie AI.

2021

Nasce Dall-E

Un sistema in grado di creare immagini realistiche o alterare immagini esistenti in base alle descrizioni fornite dagli utenti.

2022

ChatGPT Beta

Il 30 novembre, OpenAI presenta la beta di ChatGPT, un modello basato su **GPT-3.5**, multilingue e in grado di adattarsi alla lingua dell'utente.

Primo libro Fake

Armmaar Reshi ha creato la storia per bambini "Alice e Sparkle" utilizzando ChatGPT e illustrato con Midjourney. Il libro è stato pubblicato su Amazon senza spese di produzione, ma rimosso poco dopo a causa di controversie sulle immagini generate.

Medicina

A dicembre 2020, **Exscientia e Sumitomo Dainippon Pharma** hanno sviluppato il primo farmaco, **DSP-1181**, utilizzando l'intelligenza artificiale per il trattamento del disturbo ossessivo compulsivo, completando il processo in 12 mesi invece dei 5 anni. Nel 2021, **AlphaFold di DeepMind** ha previsto con successo la struttura di 330.000 proteine, ampliando il database AlphaFold a oltre 200 milioni di proteine. Infine, a febbraio 2022, Insilico Medicine ha avviato gli studi clinici di Fase I per la prima molecola scoperta dall'intelligenza artificiale.

2023

Bard di Google

Inizia il test pubblico di Bard, noto come "**ChatGPT di Google**", limitato agli utenti negli USA e nel Regno Unito, senza un lancio ufficiale al pubblico generale. LaMDa è un modello di linguaggio generativo simile a Chat-GPT 3.5, sviluppato da Google utilizzando dati da oltre **1,56 trilioni di parole** estratte da dialoghi su Internet.

«MITIGARE IL RISCHIO DI ESTINZIONE DOVUTO ALL'AI DEVE ESSERE UNA PRIORITÀ A LIVELLO MONDIALE, INSIEME AD ALTRI RISCHI SU SCALA SOCIALE. COME LA PANDEMIA E LA GUERRA NUCLEARE.>>

ChatGPT4

A marzo, il Future of Life Institute, con il supporto di figure come Elon Musk, ha chiesto una **pausa di** almeno **sei mesi** nello sviluppo dell'intelligenza artificiale per esaminarne gli impatti sulla società. Nel frattempo, ChatGPT sta lavorando alla sua quinta versione.

Modelli di fondazione e GenAI: la versione di Spindox

a cura di Paolo Costa

Quali reali opportunità si presentano alle aziende con l'avvento dei modelli di fondazione (foundation models) e dell'intelligenza artificiale generativa (GenAI)? Come si determina la struttura dei costi da considerare? Quali i rischi operativi, legali ed etici? Dopo avere sperimentato i principali LLM, Spindox propone oggi un framework robusto e articolato.

Quando si parla di modelli di fondazione (foundation models) e di intelligenza artificiale generativa (GenAI), non è esagerato definire «dirompente» lo scenario davanti ai nostri occhi. Con una rapidità mai vista in passato, si stanno mettendo a punto tecnologie capaci di trasformare profondamente il modo di progettare, produrre e fare business in tutti i principali comparti dell'industria e dei servizi. Prima ancora, sta già cambiando il modo di lavorare delle persone e di contribuire al valore delle organizzazioni con le proprie capacità individuali.

Ma come devono comportarsi le aziende, in

questa fase? È il caso di avventurarsi senza indugi in acque inesplorate, perché domani sarà già tardi? O è meglio, per il momento, restare spettatori e limitarsi a capire? Certo qualcuno dirà che, in questa fase, capire è già tanto. Il ruolo di Spindox, accanto alle aziende, non può che essere quello di stare un passo avanti, sperimentare e costruire i razionali in base ai quali compiere le scelte giuste. In attesa di avere tutte le risposte, occorre formulare le domande giuste. E per noi le domande giuste sono tre:

- Quali sono le reali opportunità offerte dai modelli di fondazione e dall'intelligenza artificiale generativa?
- Volendo avviare un progetto pilota in questo ambito, come se ne determina la struttura dei costi?
- Quali rischi operativi, legali ed etici devono essere considerati?

Di seguito proviamo a illustrare lo stato dei ragionamenti che stiamo facendo in Spindox, insieme ad alcuni clienti, per poi descrivere il framework operativo messo a punto per lanciare in sicurezza i primi progetti. Ma, prima di tutto, cerchiamo ci chiarire di che cosa stiamo parlando.

Che cos'è un foundation model?

Con l'espressione foundation model (modello di fondazione) ci riferiamo a modelli addestrati su un ampio insieme di dati non etichettati che possono essere utilizzati per compiti diversi, con una messa a punto minima. Possono costituire la base per molte applicazioni del modello di intelligenza artificiale. L'aspetto saliente dei modelli di questo tipo è che, attraverso l'apprendimento auto-supervisionato e un successivo lavoro di raffinamento, possono applicare le informazioni apprese in generale a un compito specifico.
Abbiamo dunque a che fare con un machine learning molto evoluto, che rimanda al concetto di intelligenza artificiale generale. Mentre la cosiddetta «narrow AI» si concentra su un singolo compito ed è limitata da vincoli che le impediscono di risolvere problemi sconosciuti, la «general AI» è in grado di svolgere un'ampia gamma di compiti, esibendo comportamenti che simulano le capacità cognitive umane. Si tratta di una grande passo avanti nell'ambito del famoso imitation game di Alan Turing.

Che cos'è l'intelligenza artificiale generativa?

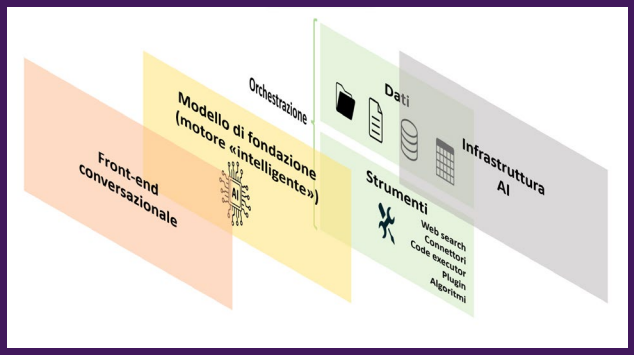
I modelli di fondazione sono il cuore dell'intelligenza artificiale generativa (GenAI), che usa reti neurali per apprendere modelli. La rete neurale viene addestrata su un ampio set di esempi e in seguito può generare nuovi dati simili al set di addestramento. In sostanza, la GenAI è un tipo di intelligenza artificiale in grado di generare nuovi contenuti, come immagini, musica, testi o ambienti virtuali, apprendendo modelli dai dati esistenti. Le architetture utilizzate sono quelle di tipo transformer (per il linguaggio naturale) e GAN (per le immagini). Altri modelli usati per la generazione delle immagini sono il Variational Autoencoder (VAE) e il Diffusion Model. Fra le tecnologie di GenAI oggi più popolari, suscitano grande interesse quelle di OpenAI (GPT, DALL-E e Whisper), ma anche Stable Diffusion di Stability, RETRO di DeepMind,

DEEPCTL di Google e LLaMA di Meta.
È bene poi precisare che non tutte le tecnologie sopra menzionate sono classificabili come modelli linguistici artificiali (Large Language Model, LLM). Per esempio, Midjourney, Stable Diffusion e DALL-E non dispongono di un modello linguistico sottostante, per cui non appartengono al sottoinsieme dei LMM. In generale, un modello linguistico artificiale è definibile come una distribuzione di probabilità su sequenze di parole: data una qualsiasi sequenza di parole di lunghezza m, un modello linguistico assegna una probabilità P all'intera sequenza.

Quando poi un LLM è capace di gestire input multimodali, ovvero di elaborare sia testi (traduzione, modellazione del linguaggio), sia immagini (rilevamento di oggetti, classificazione di immagini) sia audio (riconoscimento vocale), allora andrebbe più correttamente definito come Large Multimodal Model (LMM). È il caso di GPT-4V di OpenAI, che costituisce appunto l'evoluzione multimodale di GPT-4.0, essendo in grado di elaborare sia dati testuali sia immagini. Ad esempio, può accettare un'immagine come parte di un prompt e fornire una risposta testuale appropriata.

Non c'è solo ChatGPT

Oggi l'attenzione del grande pubblico si concentra su ChatGPT (il chatbot di OpenAI addestrato con GPT) e su Midjourney. Entrambi i sistemi sono accessibili gratuitamente o a un prezzo relativamente modesto e utilizzati per la generazione di testi e immagini. Tuttavia, al di là di queste due applicazioni, il mondo dell'intelligenza artificiale generativa è molto più vasto. E, pensando a soluzioni verticali per il business, occorre immaginare architetture complesse. Per quanto ne costituiscano il cuore, i modelli di fondazione rappresentano solo in parte di tali architetture. Uno schema generale di architettura per l'impiego della GenAI in un contesto enterprise potrebbe essere quello rappresentato nel disegno 1, che proponiamo qui sotto:



Lo strato colorato in verde corrisponde alla funzione di orchestrazione, che agisce in realtà a due livelli: da un lato tra il modello di fondazione e il front-end, dall'altro lato fra il modello stesso e le applicazioni. Il tutto deve essere poi inquadrato nel contesto di un'infrastruttura AI, come Azure OpenAI Service, Amazon Forecast o Google Cloud AI.
Particolare rilevanza hanno, in uno schema di questo tipo, funzioni come il prompt filtering (che impedisce, in entrata e in uscita, il passaggio di messaggi inappropriati), il metaprompting (ovvero la creazione di un prompt attraverso un altro prompt) e il data grounding (l'utilizzo di set di dati, documenti, reti e database esterni per fondare il modello su dati reali e sul contesto dell'utente, come spieghiamo meglio più avanti).
La gestione di tutto ciò è facilitata oggi – o, per meglio dire, guidata – da framework e pattern architetturali. È il caso di Copilot Stack, la cui espansione verso l'AI è stata presentata da Microsoft a maggio 2023.

GenAI: quali opportunità per le aziende

Ciò che colpisce, quando si parla di intelligenza artificiale generativa, è la straordinaria velocità con cui essa procede nel cammino verso la maturità tecnologica. Tuttavia, Gartner ci insegna che la maturità tecnologica non conduce in modo deterministico al plateau della produttività. È dunque importante distinguere fra prospettive di medio-lungo termine, più o meno fantasiose, e opportunità di breve termine. Ogni ragionamento deve essere supportato da un business case robusto e misurabile.
Spindox ritiene che le opportunità per le aziende legate alla GenAI si concentrino oggi in tre aree legate all'ambito del Natural Language Processing (NLP):

- Efficientamento del processo di software development (design, coding e testing)
- Sviluppo di applicazioni verticali (chatbot e voicebot, sentiment analysis, ricerca all'interno di corpora testuali, parafrasi di documenti esistenti o generazione di testi originali)
- Supporto al project management (produzione di presentazioni, report ecc.)

È verso queste tre aree che abbiamo indirizzato i nostri sforzi di ricerca per realizzare primi progetti-pilota, anche in collaborazione con quei clienti che hanno deciso di muoversi subito. In parallelo, riconosciamo l'esistenza di alcune sfide da indirizzare, alle quali corrispondono altrettanti fattori critici di successo.

Prima sfida: accrescere l'affidabilità dell'output

La prima sfida riguarda l'esigenza di garantire un'adeguata qualità dell'output. Quando usiamo ChatGPT, siamo avvisati da OpenAI che le informazioni fornite potrebbero non essere complete o accurate e che la stessa OpenAI non si assume la responsabilità di verificarne la veridicità. Com'è ovvio, una cosa simile sarebbe inaccettabile in ambito enterprise. Per questo occorre che, quando si realizzano applicazioni verticali pensate per risolvere specifici problemi di business, i modelli di fondazione siano parte di un processo di data grounding, ossia nutriti con dati di qualità, accurati e specifici, non compresi nel dataset impiegato nella fase di training degli stessi modelli. Il rischio delle cosiddette «allucinazioni» dell'algoritmo è sempre in agguato.
L'approccio RAG (Retrieval Augmented Generation) è quello oggi più riconosciuto. Si tratta di un metodo che combina un componente di ricerca («information retrieval») con un LLM. Per formulare le risposte al prompt, il modello linguistico accede a una fonte esterna di conoscenza, più ampia di quella su cui è stato addestrato e contestualizzata in termini cronologici (pensiamo, per esempio, che GPT si ferma al 2021) o di dominio (GPT è generalista). RAG riceve un input e recupera un insieme di documenti pertinenti. Questi sono concatenati con la richiesta originale di input e inviati al generatore di testo che produce l'output finale.
L'approccio RAG è implementato in Google Search Generation Experience, che integra due modelli linguistici «retrieval-enhanced» (REALM della stessa Google e il già

citato RETRO di DeepMind) con il sistema di retro-fitting RARR. Ma uno schema simile è adottato da Microsoft, che ha integrato GPT con il suo motore di ricerca Bing. C'è poi il caso del plugin per ChatGPT realizzato da AI-PRM, che dà accesso a una libreria di modelli di prompt predefiniti e selezionati.

Una strategia complementare di data grounding consiste nell'aggiunta di una prospettiva deterministà all'approccio squisitamente probabilistico della rete neurale. In quest'ottica Spindox ritiene particolarmente interessante il modello LMQL, messo a punto da ETH Zürich (ne abbiamo parlato in LMQL: dal prompt engineering al language model programming, 14 luglio 2023).

Seconda sfida: ottimizzare il costo dell'hardware

Non meno rilevante è la seconda sfida, che riguarda il costo delle risorse hardware e la loro ottimizzazione. L'intelligenza artificiale generativa richiede grandi capacità di calcolo e quindi rischia di non essere sostenibile, sia sul piano economico sia dal punto di vista ambientale. Del problema ci siamo occupati recentemente in un altro articolo sul blog di Spindox (**L'intelligenza artificiale ha fame di energia**, 14 ottobre 2023). A parte lo sviluppo di chip di nuova concezione – come le tensor processing unit (TPU), sulle quali un vendor come Google sta puntando in alternativa alle più tradizionali GPU – è importante lavorare sull'ottimizzazione dei modelli stessi. L'obiettivo è ridurre al minimo il numero di parametri, senza perdere in accuratezza dell'output.

Peraltro, i costi computazionali dei grandi modelli linguistici artificiali si riflettono sui prezzi praticati dai fornitori «as-a-service» di questo tipo di tecnologia. Sappiamo, per dire, che OpenAI pratica un modello di pricing per token (750 parole corrispondono più o meno a 1.000 token). Per l'utilizzo di GPT-4.0, la tariffa attuale è pari a 0,03 dollari per 1.000 token. Sembra un'inezia, eppure basta poco per ritrovarsi con una bolletta di migliaia di dollari al mese. Soprattutto, si tratta di un costo che può subire una brusca impen-

nata, a fronte di un significativo incremento nell'utilizzo del sistema. Possono essere valutati scenari alternativi, basati sull'hosting del LLM in un cloud pubblico, che di volta in volta si rivelano più o meno convenienti. Per esempio, l'hosting di LLaMA-2 7B nell'infrastruttura di AWS, che richiede almeno un'istanza di EC2 g5.2xlarge, nonché l'impiego di AWS API Gateway e AWS Lambda, può essere più economico rispetto a GPT-4.0. A patto che le performance di LLaMA-2 7B, che ha «solo» sette miliardi di parametri, siano sufficienti rispetto al caso d'uso che dobbiamo gestire. Perché, se viceversa optiamo per LLaMA-2 13B (tredici miliardi di parametri), i costi di hosting rischiano di triplicare.

Terza sfida: la protezione dei dati

Vi è infine una questione di natura non tecnica, ma di grande rilevanza, che riguarda la tutela dei dati personali o sensibili. L'uso dei modelli di fondazione, e in particolare dei LLM, implica in genere l'integrazione fra tecnologie open source e soluzioni commerciali. Le conseguenze, dal punto di vista della protezione dei dati, sono evidenti. Il caso d'uso più facile da immaginare è quello in cui l'organizzazione dà in pasto al modello «aperto», magari accessibile via API, informazioni che hanno natura riservata: dati dell'azienda, dei propri dipendenti o del cliente. Ovviamente si tratta di uno scenario da evitare in ogni modo, anche attraverso la definizione di specifiche policy e la formazione di tutte le persone coinvolte.

Che cosa sta facendo Spindox nell'ambito della GenAI

Da anni il nostro gruppo lavora nel campo delle tecnologie per il trattamento del linguaggio naturale (NLP), per cui oggi può contare su un'esperienza molto significativa, maturata in ambiti applicativi eterogenei, da applicare al mondo dei modelli di fondazione e dei LLM. Spindox ha così messo a punto un framework ad hoc, che lavora su sei componenti principali: 1. pipeline di data processing (gestione del flusso di dati e indicizzazione); 2. memoria (miglioramento delle performance mediante l'uso di strategie RAG, database vettoriali, knowledge graph ecc.); 3. raffinamento del LLM; 4. LLM-Ops (standardizzazione delle operazioni in produzione, compresa la definizione di template di prompt engineering); 5. orchestrazione multi-agente (per la soluzione di compiti che richiedono l'esecuzione di più agenti LLM-based); 6. API gateway (integrazione con le fonti dati e disegno dell'interfaccia utente).
Il tutto è schematizzato nella figura 2, qui sotto:

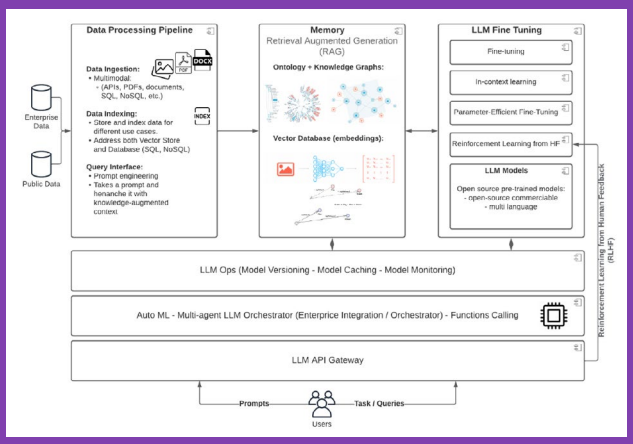


Figura 2: Framework Spindox per l'industrializzazione dei LLM.

Grazie a questo framework robusto e articolato, oggi Spindox è in grado di:

- impiegare tutti i principali LLM, commerciali e open-source, valutando pregi e difetti di ciascuna soluzione;
- implementare e industrializzare processi abilitati da LLM, sia on premise sia in modalità a servizio, garantendo la privacy e la protezione intellettuale dei dati del cliente;
- definire modelli ad hoc per personalizzare il prompt engineering in funzione del contesto del cliente, di compiti specifici da eseguire e di dati particolari che devono essere processati;
- gestire la memoria del sistema e la base di conoscenza in modo strutturato, in modo da superare i limiti degli LLM in termini di numero di token e di efficienza computazionale, a tutto vantaggio dei costi;
- costruire flussi decisionali e operativi complessi, nei quali diversi agenti basati su LLM operano contemporaneamente e collaborano per il raggiungimento di obiettivi condivisi;
- lavorare con modelli multimodali, che integrano il trattamento del linguaggio naturale con quello di contenuti di altro tipo (immagini, filmati, audio);
- superare il problema delle allucinazioni degli LLM pre-addestrati, attraverso l'approccio GAM e l'impiego di LMQL precedentemente descritti.

Approfondimenti

Simon Attard, Grounding Generative AI, Medium, marzo 2022: https://medium.com/@simon_attard/grounding-large-language-models-generative-ai-526bc4404c28.

Sebastian Borgeaud e altri, Improving language models by retrieving from trillions of tokens, Arxiv, 8 dicembre 2021 (v1) - 7 febbraio 2022 v3: https://arxiv.org/abs/2112.04426.

Luyo Gao e altri, RARR: Researching and Revising What Language Models Say, Using Language Models, Arxiv, 17 ottobre 2022 (v1) - 31 maggio 2023 (v3): https://arxiv.org/abs/2210.08726.

Kelvin Guu, REALM: Retrieval-Augmented Language Model Pre-Training, Arxiv, 10 febbraio 2020: https://arxiv.org/abs/2002.08909.

Con i fondi del PNRR nasce WICO: la piattaforma intelligente per il monitoraggio dell'acqua

Il progetto WICO, che vede Spindox come partner tecnologico per l'intelligenza artificiale, riceve il via libera formale da parte del Ministero delle Imprese e del Made in Italy (MIMIT) per l'accesso ai fondi previsti dal PNRR. Tale contributo servirà a finanziare lo sviluppo di una piattaforma di supporto alle decisioni con largo impiego di tecniche di intelligenza artificiale. Spindox svilupperà una piattaforma intelligente per il monitoraggio delle acque destinate al consumo umano. Il sistema sarà destinato all'acquedotto del Basso Valdarno, che serve oltre 800mila cittadini.

Una piattaforma intelligente per il monitoraggio dell'acqua

Il progetto WICO (Water Quality Innovative COntrol by Artificial Intelligence) si pone l'obiettivo di creare una piattaforma intelligente di monitoraggio delle acque destinate al consumo umano, che riesca a indicare in tempo reale l'insorgere di eventuali anomalie e possa funzionare da sistema di allarme rapido in caso di insorgenza di situazioni non previste (ad esempio causate da guasti, incidenti, rotture ecc.), nell'ottica di garantire la massima sicurezza per gli utenti consumatori dell'acqua destinata al consumo umano e una gestione ottimizzata da parte dei gestori del servizio idrico.

La piattaforma intelligente di monitoraggio si comporrà di due elementi principali: una rete diffusa di sensori multiparametrici installata lungo l'acquedotto, in grado di fornire in tempo reale un'impronta della qualità delle acque destinate al consumo umano; un sistema di intelligenza artificiale in grado di analizzare in tempo reali le informazioni provenienti dalla rete diffusa di sensori per interpretarne i segnali in maniera globale e unitaria e segnalare precocemente eventuali situazioni anomale. Il sistema sarà validato con il contributo di un'azienda che gestisce il servizio idrico integrato in Regione Toscana per una popolazione di oltre settecentomila abitanti.

Il progetto WICO, iniziato ad aprile 2023, è attualmente in corso e si prevede di terminarlo entro marzo 2026. Il gruppo di ricerca Spindox principalmente coinvolto elle attività è quello che fa capo ad aHead Research, il team che sviluppa e realizza soluzioni basate su Intelligenza Artificiale, sia per progetti interni che per partner e clienti. Spindox svilupperà una piattaforma di Continous Intelligence per il monitoraggio delle acque destinate al consumo umano e la rilevazione di Early Warning, in caso di insorgenza di situazioni non previste (es., causa guasti, incidenti, rotture, ecc.) In particolare, la piattaforma si occuperà di raccogliere i dati misurati dai sensori e abilitare l'esecuzione di algoritmi di calcolo basati su Artificial Intelligence rendendo consultabili i risultati tramite una web user interface che supporti gli utenti nell'identificare anomalie presenti o previste.

Gli algoritmi di Artificial Intelligence permetteranno l'analisi batch e in tempo reale dei segnali temporali multidimensionali (es., serie storiche multi-sensore, dati fotogrammetrici, immagini) e supporteranno il processo di monitoraggio real-time delle acque destinate al consumo umano, sfruttando tecniche di anomaly/outlier detection per l'identificazione preventiva dell'insorgere di situazioni di pericolo (Early Warning System). In particolare, saranno impiegati algoritmi per la previsione di serie storiche, analisi di correlazione tra segnali multidimensionali (es., serie storiche correlate a dati pro-

> **L'analisi di correlazione di segnali multimensionali sarà supportata dall'uso di reti neurali**

venienti da sensoristica spettrofotometrica), change point analysis (es., punti di cambiamento nel comportamento di un segnale) e identificazione di slow trend changes.

L'analisi di correlazione di segnali multidimensionali sarà supportata dall'uso di reti neurali innovative come le Temporal Convolutional Networks che si sono rivelate estremamente efficaci in questo tipo di contesto. Inoltre, questi modelli avanzati saranno supportati, seguendo un approccio multi-modello (come, ad esempio, strategie ensemble), da modelli come Random Forest, Support Vector Machine, Vanilla Neural Networks e Decision Trees. L'uso di questi modelli o di una combinazione di essi è in grado di estrarre nuove dipendenze altamente complesse e non lineari tra le variabili, permettendo di sfruttare l'integrazione di dati provenienti da un ventaglio ampio ed eterogeneo di sensori.

Lo sviluppo della piattaforma seguirà il paradigma ASD (Agile Software Development) in modo tale da sviluppare subito un "proof of concept prototype" che possa evolvere nel tempo con rilasci continui in modo tale da i) poter percepire in fase già esecutiva la bontà di quanto verrà progettato successivamente e ii) per interagire con le attività gestite dagli altri partner di progetto. In particolare, gli artefatti - prototipi iniziali – verranno condivisi ad ogni iterazione di sviluppo attraverso una piattaforma di collaborazione che opererà per allineare ad ogni iterazione i risultati intermedi all'intero sviluppo, garantendo un'integrazione efficace ed efficiente dei moduli di IA all'interno della piattaforma.

Il grado di innovazione del sistema che si sta sviluppando è molto elevato e contribuirà a rendere positivamente distinguibili i nuovi servizi che Spindox si propone di mettere a punto per poterli successivamente vendere sul mercato. L'architettura logica permetterà di implementare il sistema di monitoraggio delle acque permettendo il facile inserimento di moduli di IA garantendo il rilevamento di anomalie e di previsioni in tempo reale su dati di serie temporali provenienti da sensori in un contesto Big-Data Analysis. L'insieme della piattaforma di Continous Intelligence con la rete diffusa multi-sensore si declina per essere un nuovo prodotto per il mercato. Ad oggi, l'offerta Spindox include la vendita di prodotti e/o servizi SaaS basati su questi prodotti, che operano singolarmente o in modo integrato ma mai in forma così strutturata e supportata da sensoristica.

Con questo progetto si apriranno nuovi spazi su mercati attualmente poco presidiati, con l'obiettivo di raggiungere elevati volumi proponendo un approccio innovativo e particolarmente performante.

I partner del progetto WICO

Il progetto è ammesso formalmente dal Ministero delle Imprese e del Made in Italy ai finanziamenti previsti dal PNRR nell'ambito della Missione 4 (Istruzione e ricerca), Componente 2 (Dalla ricerca all'impresa). Il progetto (protocollo n. 227, Avviso Pubblico Accordi per l'Innovazione DM 31/12/2021 – I sportello), vede Spindox nel ruolo di partner tecnologico e coinvolge altri quattro attori: IBF-CNR, Archa, Acque, Dielectrik. Il totale dell'investimento dei cinque partner è pari a 5.929.642,50 euro, cui corrisponde un'agevolazione totale di 2.848.716,12 euro, di cui 1.075.079,06 euro a fondo perduto per Spindox.

Spindox lavorerà con la sede di Pisa dell'Istituto di Biofisica del CNR (Consiglio Nazionale delle Ricerche), il centro di ricerca privato e capofila del progetto Archa, il gestore della rete idrica in Toscana Acque e l'azienda di sviluppo della sensoristica Dielectrik.

- ARCHA (www.archa.it): capofila del progetto, centro ricerca privato che svolge attività di ricerca industriale e di consulenza tecnologica per l'innovazione delle imprese.
- Dielectrik (www.dielectrik.it): azienda che sviluppa e produce elettronica innovativa di diverso genere e in vari settori della ricerca e della tecnica.
- Acque (www.acque.net): azienda che gestisce il servizio idrico integrato del Basso Valdarno, che comprende una popolazione di circa 800.000 abitanti in 55 comuni delle province di Firenze, Lucca, Pisa, Pistoia, e Siena.
- CNR-Istituto di Biofisica (www.ibf.cnr.it): L'Istituto di Biofisica è uno degli Istituti del Dipartimento di Scienze fisiche e tecnologie della materia del Consiglio Nazionale delle Ricerche. Riunisce biochimici, biologi, chimici, fisici e matematici; il suo punto di forza è rappresentato dalle competenze multidisciplinare e dall'approccio interdisciplinare alle varie tematiche di ricerca che coprono argomenti complementari ad ampio raggio.

06

REXASI-PRO compie un anno

> **Il progetto mira a rilasciare un quadro ingegneristico per dispositivi mobili destinati allo spostamento autonomo di persone dall ridotta capacità motoria**

Cronaca ravvicinata di un anno di ricerca

Uno, due, tre use case. Dieci partner di progetto da sei differenti Paesi. Dodici mesi di lavoro. Sono alcuni dei numeri di REXASI-PRO, progetto triennale il cui acronimo sta per REliable & eXplAinable Swarm Intelligence for People with Reduced mObility.
Si è tenuta a Trento gli scorsi 9 e 10 ottobre la prima General Assembly. Trenta i convenuti in rappresentanza di partner chiamati al confronto interno sullo stato di avanzamento delle attività ad un anno dal loro avvio. Oltre a un workshop tecnico su strumenti e metodologie in uso in ambito progettuale, l'ordine dei lavori ha previsto anche una Light Review da parte di revisori esterni.
REXASI-PRO è infatti stato finanziato dalla Commissione Europea nell'ambito del programma Horizon Europe – Research and Innovation Actions. Il progetto mira a rilasciare un quadro ingegneristico per dispositivi mobili destinati allo spostamento autonomo di persone dalla ridotta capacità motoria.

Use case e stato dell'arte

La messa a punto di un sistema di navigazione sociale basato su veicoli guidati da Intelligenza Artificiale ha previsto nel corso del primo anno la definizione di tre casi d'uso:

- Sedie a rotelle dotate di sensori di bordo, le quali, su input dell'utente, si muovono, al chiuso o all'aperto, in ambienti più o meno affollati.
- Droni deputati allo studio degli ambienti e alla loro scansione per poter generare mappe 3D e digital twin a fini di simulazione.
- Un sistema di navigazione collaborativa risultante dalle informazioni dei precedenti casi d'uso che sia in grado di assicurare un'esperienza confortevole e sicura, anche in situazioni di emergenza.

Al giro di boa dei dodici mesi, lo sviluppo del primo use case è da considerarsi ben avviato. Ciò in considerazione della parte hardware e software del driving assistant, della robustezza del people tracker, degli esperimenti di navigazione condotti in ambienti chiusi. Per il secondo use case gli approfondimenti in corso sono volti, per un verso, ad ottimizzare l'esplorazione degli spazi dal punto di vista del tempo impiegato

dai droni e del relativo consumo di energia. Per altro verso, sono incentrati sulla ingegnerizzazione della tecnologia, in funzione di una sua accresciuta accettabilità sociale, con droni meno voluminosi e provvisti di una carenatura più leggera.

Intelligenza Artificiale, etica e comunità scientifica

Nella stessa direzione si è posta l'elaborazione di un quadro di riferimento per la valutazione delle questioni etiche correlate allo sviluppo del sistema, in tutte le sue fasi. Un sistema chiamato a porre al centro l'esperienza dell'utente finale, considerando anche l'eventualità che questi appartenga a gruppi potenzialmente vulnerabili o a rischio di esclusione.
L'apertura alla comunità scientifica ha accompagnato le fasi di ricerca sin da subito. Lo attesta la partecipazione di membri del team a conferenze e forum internazionali in Europa (Siviglia, Salonicco, Lisbona, Santiago de Compostela, Utrecht) e Stati Uniti (San Antonio). O, ancora, il coinvolgimento degli studenti dei corsi universitari attraverso forme laboratoriali. Buone pratiche di confronto allargato inaugurate nel marzo scorso dal Clustering Workshop Establishing the next level of "intelligence" and autonomy, che ha visto protagonisti i progetti rientrati nel programma Horizon Europe attraverso la medesima misura di finanziamento: A Human Centred and Ethical Development of Digital and Industrial Technologies.

Comunicare la trustworthiness

La sezione news del sito web si è fatta strumento di una comunicazione aggiornata sull'andamento della ricerca nel corso dei mesi. Non solo perché richiesto dagli adempimenti di progetto. O perché, ancora, il nodo del conciliare il know-how dei technology provider coinvolti con le esigenze dettate dalla trustworthiness si conferma la più progressiva delle sfide. Ma perché proprio questo tema, quello della trustworthiness, resta fondativo in Unione Europea per qualsiasi sistema di AI riconoscibile come legalmente affidabile, aderente a principi etici non derogabili, capace nel proprio sviluppo di dimostrarsi tanto robusto da poter escludere il verificarsi di danni anche non intenzionali. La cura nella comunicazione di questo aspetto è da considerarsi centrale nel caso di REXASI-PRO e lo sarà sempre più per Spindox.

07

Accessibili si nasce... ma si può anche diventare

Da prerogativa di sviluppo, l'accessibilità web è scesa progressivamente in secondo piano rispetto alla ricerca dell'estetica. Come si può tornare in contatto con la tematica, nel contesto moderno?

Accessibilità Web: ritorno alle origini

"Il potere del Web sta nella sua universalità. L'accesso per tutti, indipendentemente da disabilità, è un aspetto essenziale".

Queste sono le parole del fondatore del web, Time Berners-Lee, che, nel 1997, presenta l'accessibilità come una componente naturale del web, una prerogativa di sviluppo più che un optional secondario. La sensibilità nei confronti dell'accessibilità digitale viene, ai tempi, dimostrata anche con la fondazione della Web Accessibility Initiative o WAI, progetto specificamente dedicato a abbattere le barriere di accesso al web da parte di persone diversamente abili, promuovendo design, strumenti e linee guida per siti web più inclusivi.

Tuttavia, dopo oltre un quarto di secolo è possibile riscontrare una discrepanza tra il livello di sviluppo del web e l'implementazione dell'accessibilità. Quest'aspetto, infatti, viene purtroppo messo in secondo piano in favore di componenti estetiche di appeal, branding e design accattivanti. Il pubblico nascosto degli utenti diversamente abili viene quindi lasciato, trovandosi troppo spesso di fronte a siti difficilmente interagibili.

Fortunatamente, gli ultimi anni hanno visto un cambiamento di rotta nei confronti dell'accessibilità dei siti web. Sulla di una consapevolezza

Fortinatamente, gli ultimi anni hanno visto un cambiamento di rotta nei confronti dell'accessibilità dei siti web.

nei confronti delle criticità di inclusività del web, l'Unione Europea si è dotata di una normativa specificamente dedicata all'accessibilità digitale European Accessibility Act (2019/882). Concretamente, la direttiva ha lo scopo di fornire agli 87 milioni di persone affette da disabilità in Europa degli standard egalitari di accessibilità a beni, servizi, e tutti gli altri aspetti della vita di un individuo, tra cui appunto l'esperienza online.

Nello specifico, l'applicazione degli standard normativi alle realtà commerciali e alle PA ha permesso un ritorno alle origini in termini di ciò che i padri fondatori avevano inteso per il web: che fosse uno strumento egalitario in grado di funzionare ed essere alla portata di tutti, indipendentemente dal livello di abilità cognitiva, uditiva, visiva e motoria.

L'applicazione italiana del Digital Accessibility Act

Per vedere un'applicazione italiana della normativa europea in termini di accessibilità digitale dobbiamo aspettare il 2020, con l'emanazione delle Linee Guida sull'Accessibilità degli strumenti informatici per indirizzare la PA verso l'erogazione di servizi sempre più accessibili.

Il giro di vite normativo arriva però a maggio 2022, con la ricezione della direttiva europea e l'estensione degli oneri, oltre che alla PA, anche alle aziende di grandi dimensioni (tutte quelle con fatturato medio annuo per 3 anni maggiore a € 500 milioni).

Queste dovranno presentare annualmente un'apposita Dichiarazione di Accessibilità, il culmine di una serie di procedure:

- Verifiche di accessibilità degli strumenti informatici in uso (siti web, app, ecc.) per valutarne lo stato di conformità;
- Effettuare una "verifica soggettiva" per contratti di fornitura sopra soglia comunitaria;
- Compilare e pubblicare una "Dichiarazione di Accessibilità" (sotto responsabilità del Responsabile

per la transizione digitale – RTD);

- Predisporre un Meccanismo di Feedback per consentire a utenti e cittadini di inviare una segnalazione.

L'applicazione della normativa in tema di accessibilità sarà intuitivamente un'attività di profonda e impegnativa trasformazione sociale, adesso come nel prossimo futuro: a partire dal 2025, infatti, l'adeguamento interesserà tutti gli attori economici con risorse digitali.

In questo contesto, le aziende più previdenti scelgono di intraprendere un percorso verso l'implementazione accessibile dei propri siti web anche prima di essere costretta dal vincolo normativo.

Gli standard di accessibilità oggi

Quali sono quindi gli standard di accessibilità adatti a rendere il Web una risorsa fruibile in maniera egalitaria? Abbiamo precedentemente menzionato la WAI, l'iniziativa specifica sull'accessibilità web. L'organizzazione che la lancia, il World Wide Web Consortium (o W3C) fa ancora oggi da spina dorsale allo sviluppo del web. È infatti grazie al W3C che abbiamo anche le linee guida sull'accessibilità, o WCAG. Nella loro versione più aggiornata, le WCAG distinguono tre livelli di accessibilità:

- Level A – Accessibilità di base: appartengono a questa categoria tutte le specifiche fondamentali e necessarie per un corretto funzionamento di base del sito. La non conformità a questi 30 criteri può generare gravi problematiche di accessibilità;
- Level AA – Accessibilità alta: questo livello aggiunge 20 criteri a quelli già espressi dal Level A, con l'obiettivo di rendere i contenuti web accessibili in una serie più ampia di contesti;
- Level AAA – Accessibilità perfetta: Quest'ultimo aggiunge 28 criteri ai 50 già menzionati a livello A e AA, per un totale di 78 accorgimenti per rimuovere tutte le criticità relative all'accessibilità del sito web.

Dato che questi livelli identificano un crescendo in termini di compliance con gli standard di accessibilità, ce li possiamo immaginare come tre insiemi, dove A e AA sono sottoclassi di AAA.

All'aumentare del livello di sofisticatezza aumentano anche i criteri da rispettare, per cui ogni livello è interamente contenuto nel livello successivo. L'obiettivo dovrebbe essere quello di rendere un sito web il più accessibile possibile, ma lo scenario descritto nel Level AAA – Accessibilità perfetta è difficile da realizzare a causa della creazione di interazioni e strutture che contraddicono i 78 criteri di accessibilità perfetta. Pertanto, le normative sull'accessibilità digitale hanno identificato i criteri AA come quelli che devono essere rispettati. Questo vale anche per la normativa europea.

Cogliere l'accessibilità nel 2023

Comprendere l'accessibilità di un prodotto o servizio, digitale come non, nel 2023 non significa soffermarsi a banali questioni di leggibilità, contrasto, e testo alternativo. I prodotti e servizi accessibili devono avere un design sì

> **I prodotti e servizi accessibili devono accompagnare l'utente con disabilità lungo tutto il processo di interazione**

di impatto, ma anche in grado di accompagnare l'utente con disabilità lungo tutto il processo di interazione, senza ambiguità e con l'obiettivo di offrire la stessa esperienza riservata ad altri utenti.

Questo rende necessaria una riflessione approfondita di come rendere il design capace di rispondere ai diversi tipi di abilità e disabilità con cui si può interfacciare. Questo comprende:

- Disabilità visive (cecità, ipovisione, daltonismo, ecc.): per permettere un corretto funzionamento dello screen reader (software che legge il testo digitale ad alta voce) il design fornire informazioni uniche sulla posizione e ordine di lettura del testo sulla pagina. Inoltre, è importante fornire opzioni aggiuntive quali l'ingrandimento del testo;
- Disabilità uditive (DHH): per rispondere alle necessità degli utenti con gravi livelli di perdita uditiva, è opportuno fornire alternative testuali a tracce e inserti audio, sottotitolare i contenuti video (e farlo manualmente per garantire la massima precisione) e fornire semplificazioni testuali;
- Disabilità motoria (per cause fisiche e/o neurologiche): per agevolare utenti con disabilità motorie è fondamentale fornire ampi spazi web facilmente navigabili e bottoni e pulsanti facilmente cliccabili e che non richiedano particolare precisione di puntamento. È inoltre utile lasciare spazio tra i campi dei moduli, senza raggrupparli e/o fornendo scorciatoie di compilazione e scorrimento;
- Deterioramento cognitivo (depressione, schizofrenia, dislessia, ADD): per rendere l'esperienza web accessibile a questo tipo di pubblico è suggeribile fornire i contenuti facili da comprendere e disponibili in diversi formati, concentrando l'attenzione sui contenuti importanti e riducendo al minimo contenuti inutili e pubblicità.

Come avviare il percorso verso l'accessibilità

Una volta compreso il contesto normativo, esplorato i requisiti e i livelli di implementazione dell'accessibilità, ed esplorato le sfaccettature delle direzioni in cui la nostra accessibilità deve muoversi, è giunto il momento di muovere i primi passi concreti.
Dato l'alto valore che la conformazione agli standard di accessibilità apporta ad aziende ed enti economici online, la metodologia da applicare dipende dal contesto in cui quest'operazione viene inserita. A scenari di-

versi corrispondono strategie diverse affinché la transizione accessibile sia omogenea e strutturata.

Prendiamo l'esempio di siti web di piccole dimensioni. Blog di freelancer, landing page di prodotti, siti di piccole realtà commerciali e imprenditoriali. Queste sono realtà in cui, probabilmente, non esiste un reparto o figura specifica incaricata del Marketing e della comunicazione. In questi casi, le criticità di accessibilità sono facilmente correggibili attraverso l'applicazione delle Linee Guida della normativa. Una valida aggiunta (ma non alternativa) è anche la frequentazione di forum e blog online, come il sito del W3C, le linee guida WCAG, A11y, e così via.
Un ultimo step chiave, in questo contesto di dimensioni ridotte, è sicuramente la consultazione della letteratura e l'assorbimento delle best practices per quanto riguarda la creazione di contenuti in futuro.

Spostandoci su siti web di dimensioni medio- grandi, lo scenario tende a complicarsi e richiedere strategie più strutturate. Questo è il caso di aziende medio-grandi e siti che iniziano ad essere dati e o complessi, dove una o più persone possono o devono lavorare alla struttura del sito. In questi casi l'alleato migliore è sicuramente uno strumento diagnostico, perché fornisce report empirici, ripetibili e spiegabili sulle attuali criticità e indicazioni su come risolverle. Su internet, gli strumenti diagnostici sono disponibili in abbondanza, anche gratuitamente: i più raccomandabili sono Wave, Lighthouse e Accessibility Inspector. Durante il processo di diagnostica e correzione della struttura esistente, poi, è sempre opportuno avviare processi di formazione e awareness sulla tematica e intrecciare le best practices alla cultura aziendale, limitando le criticità future.

Infine, l'ultimo scenario ha bisogno di un effort ancor maggiore, rispetto all'applicazione di una checklist o l'utilizzo di un tool diagnostico. D'altronde, gli strumenti diagnostici non fanno altro che riportare osservazioni e suggerimenti. L'operazione di correzione è sempre manuale e sostenibile fino a un certo numero di ricorrenze. Per siti web molto antiquati o estremamente ramificati, le correzioni andrebbero distribuite tra un ampio numero di figure e team. Questo significa sottoporre un organico ancora non completamente formato sull'argomento allo sforzo di trasformare il core della struttura del sito. Vista l'entità della trasformazione, un'alternativa preferibile è quella della consulenza esterna. In questo modo, si può avviare un processo di implementazione strutturata dell'accessibilità nel paradigma del valore aziendale. Che si tratti di una trasformazione culturale o di fornire una struttura fortemente accessibile a un sito nascente, la consulenza è la scelta giusta per coloro che vogliono assicurare il massimo dell'accessibilità per i propri prodotti digitali e di design.

OVER DATA.

Un magazine di proprietà
di Spindox sui temi
dell'artificial intelligence
e della tech culture

Contact us

info@spindox.it
www.spindox.it

spindox