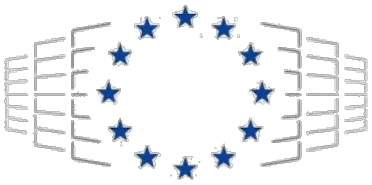
EUROCC ITALY
HPC Competence Center

WHITE PAPER

DATA LAKE

il retrobottega dell'intelligenza artificiale

a cura di: Eric Pascolo e Luca Babetto



Co-funded by the European Union. This work has received funding from the European High Performance Computing Joint Undertaking (JU) and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia under grant agreement No 101101903.

SOMMARIO

Data lake: il futuro della gestione dei dati 3
AI e Data lake, due tecnologie simbiotiche.....4
Oltre l'IA: Il Data Lake come Strumento Multiuso per l'Azienda Moderna..... 5
Data Lake su supercomputer: Velocità e scalabilità senza compromessi 6
Data lake in azione: casi d'uso nel mondo industriale.....7
Yuppies: Data Lake e HPC nel settore dell'edilizia 7
Dalla Fonte alla Risposta: Il Data Lake Ottimizza il Tuo RAG 8
Data Lake per i Digital Twin 9
Il Data Lake a un click di distanza 10

DATA LAKE: IL FUTURO DELLA GESTIONE DEI DATI

Un data lake è un sistema di archiviazione che consente di immagazzinare una grande quantità di raw data eterogenei, in formato nativo, e in maniera non strutturata. In particolare, è possibile utilizzare l'approccio "schema-on-read", in cui la struttura dei dati è definita dinamicamente al momento della lettura, a differenza del più tradizionale approccio "schema-on-write", in cui la struttura dei dati deve essere definita prima della scrittura degli stessi, richiedendo la riorganizzazione statica dell'intero database nel momento in cui sia necessario cambiarne la struttura. Questo tipo di soluzione si contraddistingue dai metodi di archiviazione più tradizionali per la sua flessibilità, versatilità, e scalabilità, permettendo non solo l'archiviazione dei dati fine a sé stessa ma anche e soprattutto l'integrazione con protocolli di analisi ed elaborazione dei dati altrimenti impossibili da effettuare.

Nel data lake i dati sono fisicamente memorizzati in una struttura piatta – senza l'utilizzo di cartelle e sottocartelle – e l'identificazione del dato viene effettuata tramite l'utilizzo di metadati. Questi sono salvati in un database rapidamente accessibile ed interrogabile tramite query, offrendo dunque la possibilità di esplorare il data lake – ad esempio filtrando per categoria di metadato o ricercando specifici tag – in maniera estremamente veloce e senza la necessità di valutare il contenuto del file se non necessario. L'interazione con il data lake è possibile in due modalità, interattiva e batch. In modalità interattiva l'utente lancia direttamente comandi verso il data lake da un terminale, con la possibilità di eseguire query e ottenere risultati in tempo reale, facilitando dunque l'analisi esplorativa e la risposta immediata a questioni specifiche e non ordinarie. In modalità batch, al contrario, è possibile richiedere l'analisi di grandi volumi di dati in blocchi, con l'opzione di programmare operazioni complesse e/o periodiche, ricevendo i risultati in un secondo momento.

Le caratteristiche più importanti del data lake sono pertanto:

- **Semplicità di raccolta dei dati**: per aggiungere contenuto al data lake, non c'è bisogno di alcun pretrattamento o strutturazione; si possono caricare i raw data in formato nativo accompagnandoli semplicemente con i metadati rilevanti.

Rapidità di elaborazione dei dati: grazie al database contenente i metadati, è possibile interrogare il contenuto dell'intero data lake in pochi secondi, selezionando i dati di interesse per le analisi in questione senza la necessità di eseguire operazioni preliminari per la ricerca.
Flessibilità e scalabilità: per espandere (o contrarre) il data lake è sufficiente aggiungere altro spazio di archiviazione! Non è necessario effettuare il processo di ristrutturazione tipica delle soluzioni storage tradizionali. Inoltre, data la natura non-strutturata di questa soluzione, è possibile memorizzare (e in seguito analizzare) dati di qualsiasi formato, anche disomogeneo tra loro.

AI E DATA LAKE, DUE TECNOLOGIE SIMBIOTICHE

L'intelligenza artificiale (IA) sta rapidamente trasformando il mondo, permeando ogni settore, dalla medicina all'industria,e rivoluzionando il modo in cui viviamo e lavoriamo. Ma silenziosamente dietro questa rivoluzione tecnologica si cela un elemento cruciale, spesso sottovalutato: i Data Lake.

Se gli algoritmi di intelligenza artificiale per migliorare devono apprendere, immaginate il data lake come la biblioteca dove gli algoritmi possono studiare, ossia un archivio digitale capace di contenere una quantità pressoché illimitata di informazioni di ogni tipo: dai semplici testi alle immagini, dai video ai dati provenienti da sensori, fino ai flussi di dati in tempo reale provenienti da dispositivi IoT, internet, post dei social network. Questa vastità e varietà di dati sempre aggiornati permetta alle IA di essere sempre più accurate e utili alle aziende.

A differenza dei tradizionali database, strutturati e rigidi, i data lake sono flessibili e aperti. Non impongono schemi predefiniti, consentendo di archiviare i dati nel loro formato originale, grezzo e non strutturato. Questa caratteristica è fondamentale per l'IA, poiché gli algoritmi di apprendimento automatico, come quelli utilizzati per il riconoscimento vocale o l'analisi predittiva, necessitano di

esplorare liberamente i dati grezzi per identificare modelli nascosti e correlazioni significative.

La flessibilità dei data lake non si limita alla struttura dei dati, ma si estende anche alla loro scalabilità, ossia la loro proprietà di immagazzinare dati mantendo le performance e questo è altro punto fondamentale per l'allenamento di algoritmi sempre più capaci di svolgere nuove operazioni in maniera accurata.

Ma i data lake non sono solo grandi contenitori di dati. Grazie a tecnologie avanzate di elaborazione parallela e distribuita, consentono un accesso rapido ed efficiente ai dati, permettendo all'IA, usata in inferenza, di analizzarli in tempo reale e di fornire risposte immediate.

In conclusione, i data lake non sono solo grandi archivi di dati, sono strumenti essenziali per lo sviluppo di IA. Possiamo affermare quindi per ogni applicazione di IA, nel retrobottega c'è un Data Lake che ne permette la realizzazione.

OLTRE L'AI: IL DATA LAKE COME STRUMENTO MULTIUSO PER L'AZIENDA MODERNA

Nell'ambito industriale, i data lake si rivelano strumenti preziosi sia per l'Operational Technology (OT) che per R&D, inclusa la prototipazione virtuale.

Nell'ambito dell'Operational Technology (OT), i data lake si rivelano strumenti di enorme valore, consentendo la raccolta e l'analisi centralizzata di una vasta gamma di dati provenienti da sensori, macchinari, sistemi di controllo e altri dispositivi presenti negli impianti industriali. Questa ricchezza di informazioni, spesso in tempo reale, permette di monitorare costantemente le performance dei processi produttivi, identificare eventuali anomalie o deviazioni dai parametri standard e intervenire tempestivamente per prevenire guasti o interruzioni.

Inoltre, l'analisi approfondita dei dati storici archiviati nel data lake consente di individuare pattern e tendenze, utili per prevedere futuri malfunzionamenti o cali

di efficienza, e pianificare interventi di manutenzione preventiva, riducendo i tempi di fermo macchina e i costi associati. L'ottimizzazione dei processi produttivi è un altro vantaggio chiave: l'analisi dei dati può evidenziare colli di bottiglia, sprechi o inefficienze, suggerendo azioni mirate per migliorare la produttività, ridurre i consumi energetici e minimizzare l'impatto ambientale.

Sul fronte della ricerca e sviluppo, i data lake offrono un ambiente ideale per l'analisi di grandi volumi di dati sperimentali e per l'integrazione di questi nelle simulazioni, accelerando la prototipazione virtuale, l'identificazione di nuove soluzioni e il miglioramento dei prodotti esistenti.

In particolare, i dati all'interno di un data lake possono rendere più accurate e significative le simulazioni, poiché si possono inserire "condizioni al contorno" reali aiutando gli algoritmi a convergere ad una soluzione migliore. Integrando questi dati storici e/o in tempo reale nelle simulazioni o applicando metodologie di analisi dati, è possibile creare modelli più accurati e realistici, validarli e calibrarli, nonché ottimizzare parametri e condizioni operative. Inoltre, l'analisi di scenari "what-if" e lo sviluppo di nuovi prodotti e processi diventano più efficienti ed efficaci grazie alla disponibilità di dati concreti.

DATA LAKE SU SUPERCOMPUTER: VELOCITÀ E SCALABILITÀ SENZA COMPROMESSI

Un data lake consente di immagazzinare quantità enormi di dati. Un supercomputer è costruito per elaborare quantità enormi di dati, il che può diventare una tecnologia abilitante per situazioni in cui il tempo sia ristretto o il problema in questione sia molto grande. Il connubio tra queste due tecnologie è pertanto naturale ed immediato. L'integrazione di un data lake come soluzione di archiviazione per dati che possono poi venire elaborati in maniera massiccia e parallela dal supercomputer apre la strada ad applicazioni che altrimenti necessiterebbero di particolari accorgimenti e know-how molto specifico.

Alcune soluzioni commerciali disponibili al giorno d'oggi consentono di processare il contenuto del data lake utilizzando risorse cloud. Questo può essere sufficiente per diverse applicazioni, in cui ciascuno dei server di calcolo lavora sul proprio contenuto in maniera autonoma e indipendente. A differenza di un cloud datacenter, in un supercomputer i server di calcolo hanno la possibilità di comunicare tra loro in maniera immediata, scambiandosi dati ed informazioni durante la fase di processamento. Questo è imprescindibile nella maggior parte dei workflow che coinvolgono intelligenza artificiale, ad esempio. Inoltre, per loro natura l'hardware di un cloud datacenter è spesso configurato pensando a generalità di utilizzo, astrazione, e scalabilità, mentre i supercomputer sono disegnati per offrire le massime prestazioni senza compromessi, sorpassando di gran lunga le capacità di un pool di risorse cloud per calcoli di grosse dimensioni.

Non da trascurare vi è anche il fatto che la maggior parte dei sistemi di supercalcolo siano a gestione pubblica, il che permette alle aziende di non dover forzatamente fare affidamento a servizi di terze parti forniti da colossi internazionali come Amazon, Google, o Microsoft.

DATA LAKE IN AZIONE: CASI D'USO NEL MONDO INDUSTRIALE

Yuppies: data lake e HPC nel settore dell'edilizia

YDMS (Yuppies Data Management System) è un progetto incentrato sulla creazione di un'infrastruttura Data Lake specificamente progettata per gestire, archiviare e organizzare dataset relativi a rilievi di edifici e altre infrastrutture civili. Per raggiungere questo obiettivo, il progetto utilizza strategicamente la potenza dei data lake e del calcolo ad alte prestazioni (HPC). YDMS è ospitato su Cloud come servizio e su HPC per la parte computazionale, l'integrazione è possibile grazie a un sistema di archiviazione duale che mappa il Parallel Filesystem a un object storage accessibile tramite l'API S3. Un data lake centralizzato consente a YDMS di archiviare diversi dataset

provenienti da rilievi, tra cui immagini, letture di sensori e report tecnici, con una flessibilità che i database tradizionali non offrono. L'integrazione di HPC fornisce la potenza di calcolo necessaria per elaborare rapidamente questi dataset complessi. Ciò consente a Yuppies di eseguire analisi sofisticate, simulazioni e modellizzazioni che sarebbero difficili o dispendiose in termini di tempo su computer standard. Combinando la scalabilità dei data lake con la potenza dell'HPC, YDMS stabilisce un'infrastruttura robusta per l'archiviazione e l'organizzazione dei dati di rilievo in un modo che ottimizza l'analisi per la gestione energetica, la gestione delle strutture e la conservazione del patrimonio pubblico nel suo complesso. Il progetto consente a Yuppies di utilizzare i dati archiviati per vari scopi, tra cui potenzialmente l'identificazione di modelli, la previsione di potenziali problemi e l'ottimizzazione della manutenzione. Il progetto ha dimostrato che l'introduzione di nuove tecniche di digitalizzazione, come l'introduzione di un Data Lake e HPC, in un campo come l'edilizia può aprire nuove possibilità in termini sia di gestione che di ottimizzazione di strutture complesse come gli edifici, portando a risparmi significativi.

Dalla Fonte alla Risposta: Il Data Lake Ottimizza il Tuo RAG

La Retrieval Augmented Generation (RAG) rappresenta un approccio innovativo nell'ambito dell'elaborazione del linguaggio naturale, che mira a superare le limitazioni dei tradizionali modelli linguistici di grandi dimensioni (LLM). La RAG combina la potenza generativa degli LLM con la capacità di recuperare informazioni pertinenti da fonti esterne, consentendo di generare risposte più accurate, contestualizzate e aggiornate.

Un elemento chiave della RAG è la scelta del database per l'archiviazione e il recupero delle informazioni. In questo contesto, l'integrazione di data lake ad alte prestazioni su supercomputer con database vettoriali (vector database) o grafici (graph database) offre vantaggi significativi.

I vector database, specializzati nella memorizzazione e nell'interrogazione efficiente di embedding ad alta dimensionalità, consentono di catturare le relazioni semantiche tra i dati, facilitando il recupero di informazioni pertinenti in

base alla somiglianza semantica. I graph database, d'altra parte, eccellono nella rappresentazione di relazioni complesse tra entità, offrendo la possibilità di esplorare connessioni profonde e complesse all'interno dei dati.

Entrambe le tipologie di database possono essere ospitate all'interno del data lake che contiene anche i dati grezzi, questo approccio ha due vantaggi, il primo la possibilità di aggiornare con l'arrivo di nuovi dati il database per la RAG, e il secondo di versionare questi DB e conservarli su un sistema scalabile, che è il supercomputer.

Il flusso di lavoro per l'aggiunta di nuovi dati è stato progettato per garantire efficienza e scalabilità. I nuovi file vengono prima aggiunti al data lake. Successivamente, uno script specializzato, eseguito sull'HPC (High-Performance Computing), dagenera nuovi embedding VDB dai file da raggare. Infine, il file contenete il database risultante viene salvato nel data lake, con metadati gestiti in modo efficiente.

La fase di "inferenza", in cui le query degli utenti vengono elaborate per generare risposte, è anche essa ottimizzata. L'utente invia una query contenente il prompt desiderato. Il data lake fornisce quindi al job su supercomputer che elabora il prompt il percorso corretto del file del database che può essere letto in parallelo. La fase di RAG viene eseguita, interfacciandosi con un modello LLM (Large Language Model) per generare la risposta finale, che viene poi restituita all'utente.

Sebbene per avere una implementazione in tempo reale bisogna assicurare una coda dedicata sullo scheduler del supercomputer, i vantaggi di questa architettura sono evidenti. Uno dei principali punti di forza è la centralizzazione di dati e calcolo in un unico ambiente, accessibile tramite una semplice interfaccia. Questa caratteristica semplifica notevolmente la gestione e l'utilizzo del sistema, offrendo un'esperienza utente intuitiva e produttiva.

Data Lake per i Digital Twin

Un Digital Twin è la rappresentazione virtuale di un sistema, processo o oggetto fisico che replica lo stato della sua controparte in tempo reale, consentendo ad esempio interventi preventivi o analisi predittive. Ad esempio, il Digital Twin di una città "intelligente" può contenere informazioni statiche come topografia, disposizione degli edifici, dati demografici, statistiche socioeconomiche, ma anche – e soprattutto – informazioni dinamiche e volatili, come traffico, stato della rete di trasporto pubblico, meteo, consumo energetico, ecc.

Tutto questo serve per creare un'immagine il quanto più completa e realistica possibile della città, il che permette di ottimizzare la pianificazione di eventuali interventi in maniera proattiva e di gestire al meglio le risorse a disposizione.

Data la natura estremamente variegata della tipologia di dati necessari per il funzionamento del Digital Twin e la necessità di aggiornare in maniera continua il suo contenuto, il Data Lake è senz'ombra di dubbio la piattaforma ideale per l'archiviazione ed il trattamento delle informazioni che costituiscono il Digital Twin, data appunto la capacità di gestione ed ingestione di enormi quantità di dati grezzi, non strutturati e di qualsiasi formato, permettendone l'accesso in maniera rapida, semplice, e soprattutto automatizzabile. Inoltre, vi è la possibilità di rendere parte del digital twin un vero e proprio servizio commerciale, in cui ad esempio viene fornita agli utenti una piattaforma per consultare alcune informazioni del digital twin per i propri interessi.

IL DATA LAKE A UN CLICK DI DISTANZA

EuroCC Italy ha reso disponibile l'applicazione Data Lake Ready to Use, atta ad equipaggiare piccole e medie imprese con la tecnologia di archiviazione e processamento di dati di cui abbiamo discusso in questo documento. Il Data Lake di EuroCC può essere personalizzato e configurato liberamente a seconda delle esigenze dell'azienda, e si integra con sistemi di supercalcolo in maniera nativa. Per l'implementazione di questo servizio, ci sono solo alcuni requisiti ottimali per sfruttare al meglio la tecnologia:

- Accesso ad una piattaforma cloud che utilizzi l'infrastruttura OpenStack, in grado di fornire la disponibilità di una macchina virtuale con almeno 8 vCPU, 32GB di RAM e 1TB di storage;
- Accesso ad un sistema HPC il cui storage sia configurato in modalità duale S3/parallel filesystem, con ovviamente un budget valido sia per quanto riguarda risorse di storage che risorse di calcolo.

Se queste infrastrutture non dovessero già essere in possesso dell'azienda, esiste comunque la possibilità di farne richiesta tramite bando europeo [https://eurohpc-ju.europa.eu/index_en], in maniera completamente finanziata da fondi pubblici.

Il deploy del servizio è stato semplificato ed automatizzato in modo da permettere il corretto e completo funzionamento di tutta l'infrastruttura del Data Lake seguendo pochi, semplici passi. Il manuale di utilizzo, contenente tutti i passaggi necessari, è disponibile a questo indirizzo [https://github.com/Eurocc-Italyhttps://github.com/Eurocc-Italy//]. Per poter effettuare l'installazione è necessario solamente un singolo PC/portatile* con accesso tramite internet alla piattaforma cloud e al sistema HPC. In seguito, sarà possibile interagire con il Data Lake da qualsiasi PC/portatile* autorizzato. L'autenticazione di utenti e dispositivi viene effettuata tramite token ed è controllata unicamente e direttamente dall'azienda. È inoltre possibile limitare l'accesso al servizio ai soli dispositivi facenti parte della rete aziendale, minimizzando così il rischio di attacchi informatici.

Per più informazioni visita www.euroccitaly.it

*per dispositivi Windows è necessario l'utilizzo di WSL