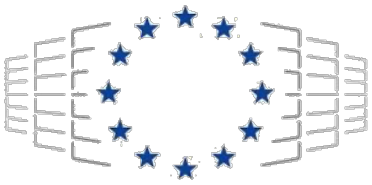
EUROCC ITALY
HPC Competence Center

WHITE PAPER

DATA LAKE
the Hidden Nook of Artificial Intelligence

Edited by: Eric Pascolo and 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.

SUMMARY

DATA LAKE: THE NEW HORIZON FOR DATA MINING

A data lake is an archival solution which allows the storage of large amounts of heterogeneous raw data, in native format, and in an unstructured fashion. In particular, it is possible to use the “schema-on-read" approach, in which the data structure is defined dynamically at the time of reading/analysis, unlike the more traditional “schema-on-write" approach in which the data structure must be defined before writing, requiring a complete and static reorganization of the whole database should a different structure be necessary.

This kind of solution differs from more traditional storage solutions for its flexibility, versatility, and scalability, allowing not only to store your data, but also (and most importantly) enabling the integration with data analysis protocols otherwise impossible to execute.

In the data lake, files are physically stored using a flat structure – without the use of folders and subfolders – and the identification of the data is carried out using metadata. These are stored in a readily accessible separate database, which allows the exploration of the data lake by the use of simple queries – for example, filtering by certain categories of metadata or selecting specific tags – in an extremely fast way without the need to evaluate the file contents if not necessary. Interaction with a data lake is possible in two modalities, interactive and batch. In interactive mode, users launch commands to the data lake directly from a terminal, with the possibility of launching queries obtaining results in real time, making it easier to perform explorative analyses and answer specific, non-ordinary questions very quickly. In batch mode, on the contrary, it is possible request the analysis of large data volumes in blocks, with the option to program complex and/or periodic operations, obtaining the results at a later time.

The most important characteristics of the data lake are therefore:

Ease of data ingestion: to add content to the data lake, there is no need for any pre-processing or structuring; you can upload files in their native format, simply by pairing them with their relevant metadata.

Rapidity of data analysis/processing: thanks to the metadata database, it is possible to query the data lake content in a manner of seconds,

selecting only the data which is relevant for the analysis/processing without the need to perform preliminary operations on irrelevant data.

Flexibility and scalability: to expand (or contract) the data lake, you only need to add more storage! There is no need to perform the restructuring operations typical of traditional storage solutions. Furthermore, given the nonstructured nature of the solution, it is possible to store (and then analyze) data of any kind and format.

AI AND DATA LAKES: A SYMBIOTIC PARTNERSHIP

Artificial Intelligence (AI) is rapidly transforming the world, permeating every sector, from medicine to industry, and revolutionizing the way we live and work. But silently behind this technological revolution lies a crucial, often underestimated element: Data Lakes.

If AI algorithms need to learn in order to improve, imagine the data lake as the library where algorithms can study, a digital archive capable of containing an almost unlimited amount of information of all kinds: from simple texts to images, from videos to data from sensors, up to real-time data streams from IoT devices, the internet, and social media posts. This vastness and variety of constantly updated data allows AI to be increasingly more accurate and useful to companies.

Unlike traditional databases, which are structured and rigid, data lakes are flexible and open. They do not impose predefined schemas, allowing data to be stored in its original, raw, and unstructured format. This feature is essential for AI, as machine learning algorithms, such as those used for speech recognition or predictive analysis, need to freely explore raw data to identify hidden patterns and significant correlations.

The flexibility of data lakes is not limited to the structure of the data but also extends to their scalability, which is their ability to store data while maintaining

performance, and this is another fundamental point for training algorithms that are increasingly capable of performing new operations accurately.

But data lakes are not just large containers of data. Thanks to advanced parallel and distributed processing technologies, they allow rapid and efficient access to data, allowing AI, used in inference, to analyze them in real time and provide immediate answers.

In conclusion, data lakes are not just large data archives, they are essential tools for the development of AI. We can therefore say that in the hidden nook of every AI application, we find a data lake.

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.

BEYOND AI: THE DATA LAKE AS A MULTI-PURPOSE TOOL FOR THE MODERN ENTERPRISE

In the industrial sector, data lakes prove to be valuable tools for both Operational Technology (OT) and R&D, including virtual prototyping.

In OT, data lakes are incredibly valuable, enabling the centralized collection and analysis of a wide range of data from sensors, machinery, control systems, and other devices present in industrial plants. This wealth of information, often in real-time, allows for continuous monitoring of production process performance, identification of any anomalies or deviations from standard parameters, and timely intervention to prevent breakdowns or interruptions.

Furthermore, in-depth analysis of historical data stored in the data lake allows for the identification of patterns and trends, useful for predicting future malfunctions or efficiency drops, and for planning preventive maintenance interventions, reducing machine downtime and associated costs. Production process optimization is another key advantage: data analysis can highlight

bottlenecks, waste, or inefficiencies, suggesting targeted actions to improve productivity, reduce energy consumption, and minimize environmental impact.

On the research and development front, data lakes provide an ideal environment for analyzing large volumes of experimental data and for integrating this data into simulations, accelerating virtual prototyping, identifying new solutions, and improving existing products.

In particular, the data within a data lake can make simulations more accurate and meaningful, as real "boundary conditions" can be incorporated, helping algorithms converge to a better solution. By integrating this historical and/or real-time data into simulations or applying data analysis methodologies, it is possible to create more accurate and realistic models, validate and calibrate them, as well as optimize parameters and operating conditions. Moreover, the analysis of "what-if" scenarios and the development of new products and processes become more efficient and effective thanks to the availability of concrete data.

DATA LAKE ON SUPERCOMPUTERS: UNCOMPROMISING SPEED AND SCALABILITY

A data lake allows users to store enormous amounts of data. A supercomputer is built to process enormous amounts of data, which can become an enabling technology whenever time is restricted or the problem at hand is very large. The union of these two technologies is therefore natural and immediate. The integration of a data lake as storage solution for data which can then be processed in a massively parallel way by a supercomputer paves the way for applications which otherwise would require particular precautions and very specific know-how.

Some commercial solutions available today allow the processing of the data lake content using cloud resources. This may be sufficient for several applications, in which each computing server works on its own content autonomously and independently. Unlike a cloud datacenter, in a supercomputer the computing servers have the ability to communicate with each other very quickly, exchanging

data and information during the processing phase. This is essential in the majority of workflows involving artificial intelligence, for example. Furthermore, by their very nature, cloud resources are configured to accommodate for general purpose computing, abstraction, and scalability, while supercomputers are designed to offer the highest performance without compromise, far surpassing the capabilities of cloud computing for large computations.

Not to be overlooked is also the fact that most supercomputing centers are state-managed, which allows companies not to be forced to rely on third party services provided by international giants such as Amazon, Google, or Microsoft.

DATA LAKE IN ACTION: USE CASES IN THE INDUSTRIAL WORLD

Yuppies: data lake and HPC in constructions sector

YDMS (Yuppies Data Management System) is a project focused on creating a Data Lake infrastructure specifically for managing, storing and organizing datasets related to surveys of buildings and other civil infrastructure. To accomplish this, the project strategically employs the power of data lakes and high-performance computing (HPC). The YDMS is hosted on Cloud as a service and on HPC for computational part, the integration is possible thanks to a dual storage system that maps the Parallel Filesystem to an object storage accessible via S3 API.

A centralized data lake enables YDMS to store diverse datasets from surveys, including images, sensor readings, and technical reports, with the flexibility traditional databases lack. The integration of HPC, provides the computational muscle needed to process these complex datasets quickly. This allows Yuppies to perform sophisticated analytics, simulations, and modelling that would be difficult or time-consuming on standard computers.

By combining the scalability of data lakes with the power of HPC, YDMS establishes a robust infrastructure for storing and organizing survey data in a way

that optimizes analysis for energy management, facilities management, and the preservation of the public heritage as a whole. The project enables Yuppies to use the stored data for various purposes, potentially including identifying patterns, predicting potential issues, and optimizing maintenance.

The project has demonstrated that introducing new digitalization techniques, such as the introduction of a Data Lake and HPC, in a field like construction can open up new possibilities in terms of both management and optimization of complex structures such as buildings, leading to significant savings.

From Source to Answer: the data lake optimizes your RAG

Retrieval Augmented Generation (RAG) represents an innovative approach in the field of Natural Language Processing (NLP), which aims to overcome the limitations of traditional Large Language Models (LLMs). RAG combines the generative power of LLMs with the ability to retrieve relevant information from external sources, allowing the generation of more accurate, contextualized, and up-to-date responses.

A key element of RAG is the choice of database for storing and retrieving information. In this context, the integration of high-performance data lakes on supercomputers with vector databases or graph databases offers significant advantages.

Vector databases, specializing in the efficient storage and querying of high-dimensional embeddings, enable the capture of semantic relationships between data, facilitating the retrieval of relevant information based on semantic similarity. Graph databases, on the other hand, excel in representing complex relationships between entities, offering the possibility to explore deep and intricate connections within the data.

Both types of databases can be hosted within the data lake that also contains the raw data. This approach has two advantages: first, the ability to update the RAG database with the arrival of new data, and second, to version these databases and store them on a scalable system, which is the supercomputer.

The workflow for adding new data has been designed to ensure efficiency and scalability. New files are first added to the data lake. Subsequently, a specialized script, executed on the HPC (High-Performance Computing), generates new VDB embeddings from the files to be raggregated. Finally, the file containing the resulting database is saved in the data lake,with metadata managed efficiently.

The "inference" phase, in which user queries are processed to generate responses, is also optimized. The user submits a query containing the desired prompt. The data lake then provides the job on the supercomputer that processes the prompt with the correct path of the database file that can be read in parallel. The RAG phase is executed, interfacing with an LLM (Large Language Model) to generate the final response, which is then returned to the user.

Although a dedicated queue on the supercomputer's scheduler is necessary to have a real-time implementation, the advantages of this architecture are evident. One of the main strengths is the centralization of data and computation in a single environment, accessible through a simple interface. This feature greatly simplifies the management and use of the system, offering an intuitive and productive user experience.

Data Lake for Digital Twin

A Digital Twin is the virtual representation of a system, process, or physical object, which replicates the status of its counterpart in real time, allowing for example preventive action or predictive analyses. For example, the Digital Twin of a smart city could contain both static information such as topography, building layout, demographics, socio-economic statistics, as well as – and maybe most importantly – dynamic and volatile information such as traffic data, public transport status, weather, energy demand, etc. All of this serves to create as complete and realistic a picture of the city as possible, which enables the optimization and planning of operations in a proactive way, and to better administer available resources.

Given the extremely varied nature of necessary data for the functioning of the Digital Twin and the need to continuously update its content, a Data Lake is

without shadow of doubt the ideal platform for storage and processing of the Digital Twin data, given the ability to handle and ingest enormous amounts of raw, unstructured data of any format, allowing rapid and simple access in an easy to automate fashion. Furthermore, there is the possibility of converting part of the digital twin into an actual commercial service, in which for instance a platform is provided to users for consulting some information of the digital twin for their own interests.

DATA LAKE: ONLY ONE CLICK AWAY

EuroCC Italy has developed the Data Lake Ready to Use application, aimed at equipping small and medium enterprises with the data storage and processing technology we have discussed in this document. The EuroCC Data Lake can be configured and personalized freely according to the company needs, and natively integrates with supercomputing systems. For the deployment of this service, there are a few prerequisites to optimally utilize this technology:

- Access to a cloud platform using the OpenStack infrastructure, able to provide a virtual machine with at least 8 vCPUs, 32GB of RAM and 1TB of storage.
- Access to a HPC system in which the storage is configured in dual S3/parallel filesystem mode, with a valid budget both in terms of storage and computing resources

Should this infrastructure not already be in the company's possession, there is also the possibility to access it via a fully funded European call [https://eurohpc-ju.europa.eu/index_en].

The deployment of the service has been simplified and automated to enable the complete and correct functionality of the whole Data Lake infrastructure by only following a few simple instructions. The user manual, which contains all the necessary steps, is available at this link [https://github.com/Eurocc-Italy/]. To carry

out the installation only a single PC/laptop* with internet access to the cloud and HPC platforms is necessary. After this, it will be possible to interact with the Data Lake from any authorized PC/laptop*. User and device authentication is carried out via tokens and is controlled uniquely and completely by the company. It is also possible to limit access to the service to only devices in the company network, minimizing risks of cyber-attacks.

For more info visit www.euroccitaly.it

*for Windows machines, WSL is required