UNIVERSITÀ POLITECNICA DELLE MARCHE

DII Dipartimento Ingegneria Informazione

L' AI per l'accesso e l'integrazione di dati in ambienti Big Data

Emanuele Storti

Knowledge Discovery & Management group



Claudia Diamantini



Domenico Potena



Emanuele Storti



Alex Mircoli



Alessandro Mele



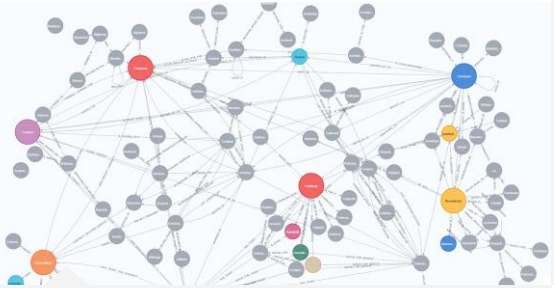
Cristina Rossetti



Monica Marconi Sciarroni (ex)

Introduzione

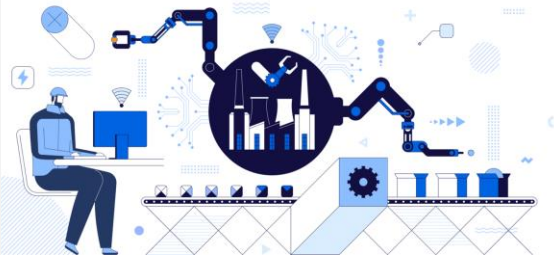
Obiettivo: supporto all'integrazione semantica, all'interrogazione, analisi e governance dei dati



Applicazioni in scenari di Big Data ed ambienti federati, Data Lake, gestione dei processi, IoT e Industria 5.0



Utilizzo di linguaggi e tecnologie semantiche per la rappresentazione della conoscenza e ragionamento ibrido (IA simbolica e IA generativa)



Data management in ambienti distribuiti

In contesti distribuiti ed autonomi, possono coesistere sorgenti dati eterogenee per

- **formato**: strutturato, semi-strutturato, …
- **schema**: diversità di attributi, tipi di dato e modelli di rappresentazione
- **semantica:** ambiguità interpretativa



Trasformare **sorgenti dati isolate** in un **ecosistema informativo interoperabile, governato e sicuro**.

Data management in ambienti distribuiti

In contesti distribuiti ed autonomi, possono coesistere sorgenti dati eterogenee per

- **formato**: strutturato, semi-strutturato, …
- **schema**: diversità di attributi, tipi di dato e modelli di rappresentazione
- **semantica:** ambiguità interpretativa

Trasformare **sorgenti dati isolate** in un **ecosistema informativo interoperabile, governato e sicuro**.

Requisiti:

- rappresentare i dati con formati uniformi ed interoperabili per supportarne l'integrazione
- descriverli con un set ricco di metadati per supportare una governance dei dati
- sistemi deterministici per estrarre informazione garantendo correttezza e consistenza

Symbolic AI (Ontologie, grafi, reasoning)

Linguaggi di rappresentazione formale della conoscenza basati su linguaggi logici e sistemi di inferenza deduttiva

Ontologie (schema)

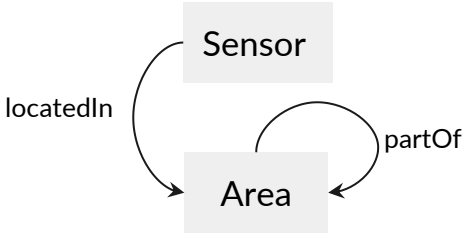
- **Modello concettuale unico** per entità e relazioni di dominio.
- **Vocabolario comune:** risolve le eterogeneità semantiche

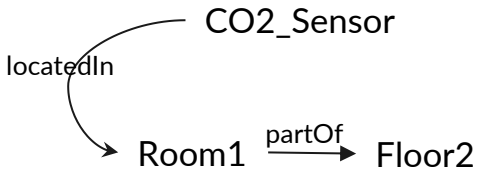
Knowledge Graph (dati)

- **Rappresentazione uniforme** dei dati
- **Interconnessione** tra dati eterogenei

Reasoning logico

- **Motore di inferenza** per derivare conoscenza a partire dai dati e dallo schema
- **Verifica di coerenza e consistenza**



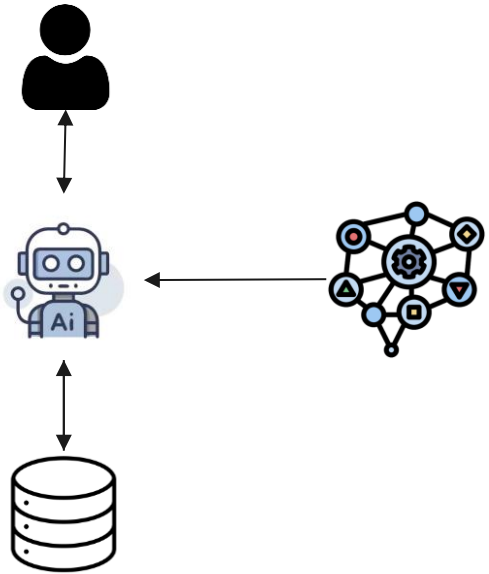


Area(x), Area(y), partOf(y,x), right(user,x) → right(user,y)

Neurosymbolic AI (LLM, agenti)

Unire la flessibilità dell'IA subsimbolica (generativa) al rigore logico dell'IA simbolica per un accesso ai dati più semplice, sicuro, accurato e minimizzando il rischio di allucinazioni grazie al grounding sul Knowledge Graph.

- **Traduzione da linguaggio naturale a query:** interpretazione dei requisiti e intenti dell'utente tramite LLM + KG.
- **Coordinamento** tramite agenti AI vincolati da regole logiche formali.
- **Verbalizzazione dei risultati:** sintesi e spiegazione chiara degli output strutturati prodotti dal sistema.
- **UX conversazionale:** accesso semplificato anche per utenti non tecnici.



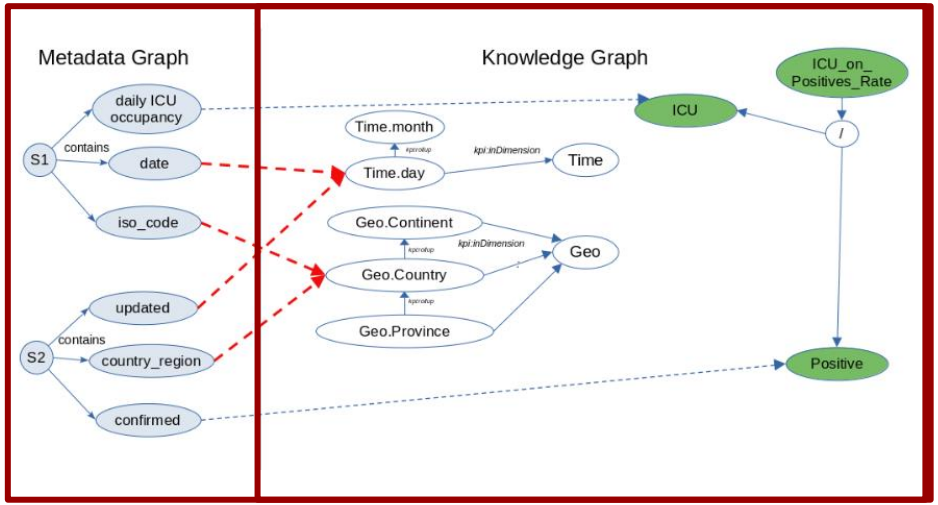
Data Lake

Problema: trovare sorgenti dati che soddisfano i requisiti dell'utente all'interno di un vasto repository di sorgenti eterogenee (prive di uno schema esplicito e documentato)

- Esempio: "*ICU by country and day*"

Rappresentazione:

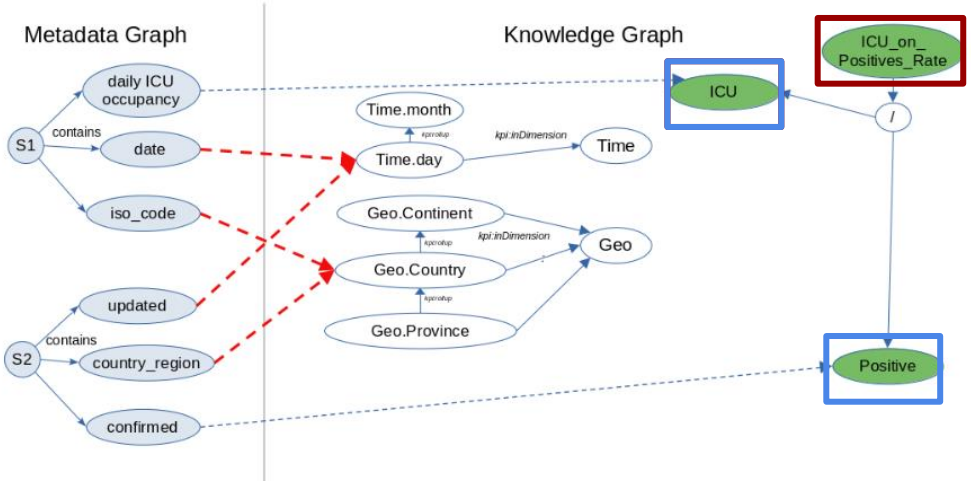
- i metadati di ogni sorgente sono descritti in un grafo
- ciascun attributo delle sorgenti viene mappato a un concetto di un **Knowledge Graph** (tramite locality-sensitive hashing)



Data Lake

Query answering:

- Il sistema individua le sorgenti dati che forniscono le informazioni richieste dall'utente sfruttando i mapping
- **Reasoning logico + reasoning algebrico:** le combina per calcolare dinamicamente il risultato tramite query reformulation



C. Diamantini, D. Potena, E. Storti, "*Analytic Processing in Data Lakes: A Semantic Query-Driven Discovery Approach*", Information Systems Frontiers, Springer, 2024

Data Lake

Applicazioni del motore di reasoning logico + algebrico: data integration, quality control, business intelligence, compliance ESG



Analisi:

- trovare **catene di dipendenze** tra indicatori



Calcolo:

- derivare **formule alternative di calcolo**
- **calcolo dinamico** di nuovi indicatori da quelli disponibili



Confronto:

- trovare **somiglianze strutturali** tra due indicatori
- trovare **dipendenze comuni** per un insieme di indicatori



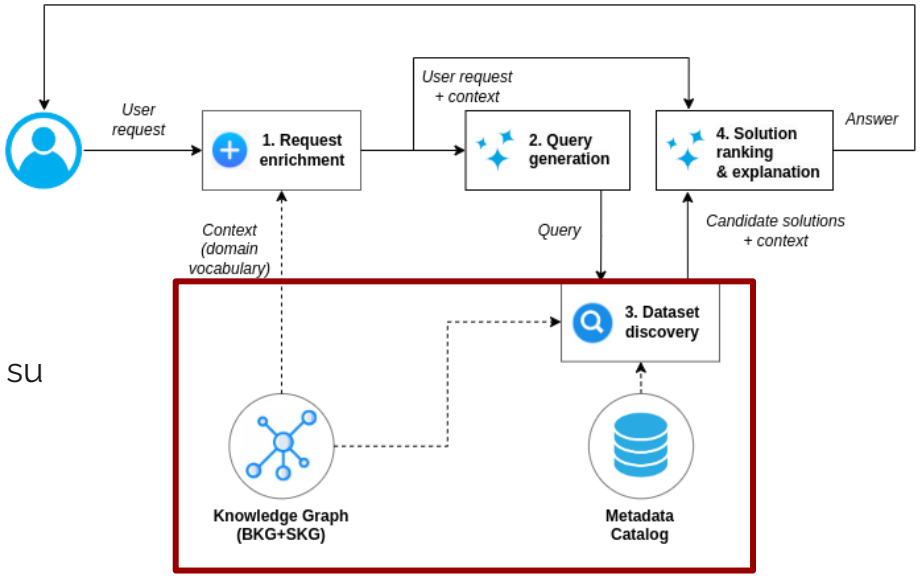
Data management

- verifica automatica di coerenza e consistenza

Data Lake: KG+LLM per explainability

Dataset discovery supportato da LLM+KG

- L'utente esprime una richiesta in linguaggio naturale
  - *"Trova dati relativi all'inquinamento nel centro Italia nell'ultimo quinquennio"*
- La richiesta viene arricchita da informazioni su misure e dimensioni
- Il LLM interpreta la richiesta e scrive la query corrispondente, che viene eseguita
- I risultati vengono di nuovo forniti al LLM per ranking e generazione della risposta

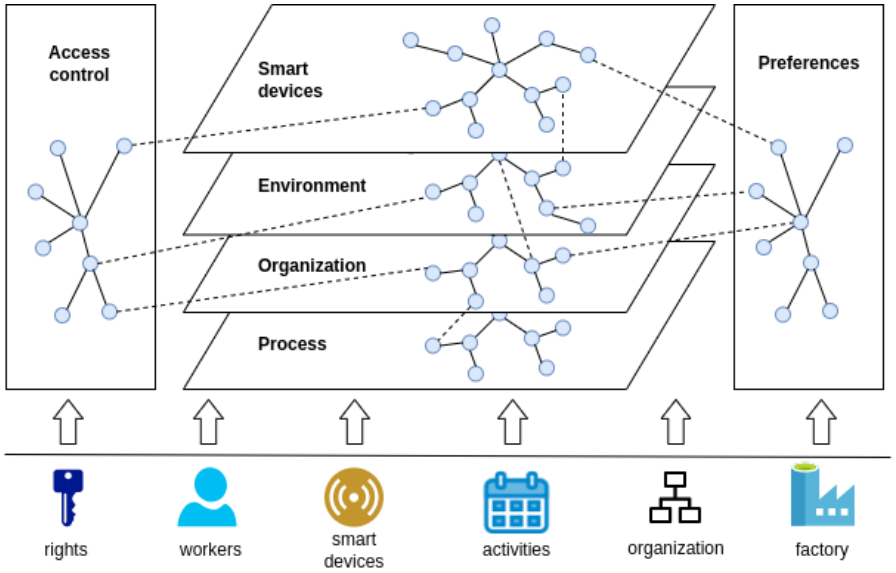


Diamantini, C., Mele, A., Mircoli, A., Potena, D., Rossetti, C., & Storti, E. A Graph RAG Approach to Enhance Explainability in Dataset Discovery. *Data Science and Engineering*, 1-23, 2025.

Industry 5.0

Un Enterprise Knowledge Graph consente di:

- rappresentare in modo uniforme la conoscenza organizzativa
- interconnettere dati ed oggetti eterogenei
- supportare query più espressive sui dati



M. Arazzi, A. Nocera, E. Storti. "The SemIoE Ontology: A Semantic Model Solution for an IoE-based Industry." *IEEE Internet of Things Journal, Elsevier*, 2024.

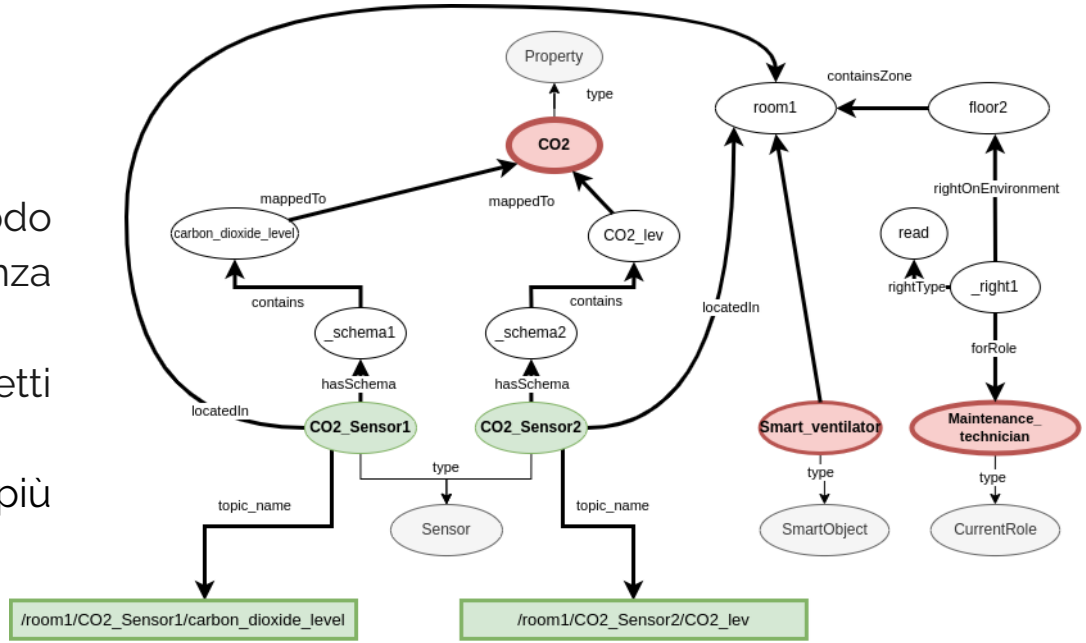
Finanziato dall'Unione europea
NextGenerationEU

Industry 5.0

Un Enterprise Knowledge Graph consente di:

- rappresentare in modo uniforme la conoscenza organizzativa
- interconnettere dati ed oggetti eterogenei
- supportare query più espressive sui dati

"Trova sensori di CO2 co-localizzati con lo Smart_ventilator, per il quale il Maintenance_technician ha permessi di lettura"



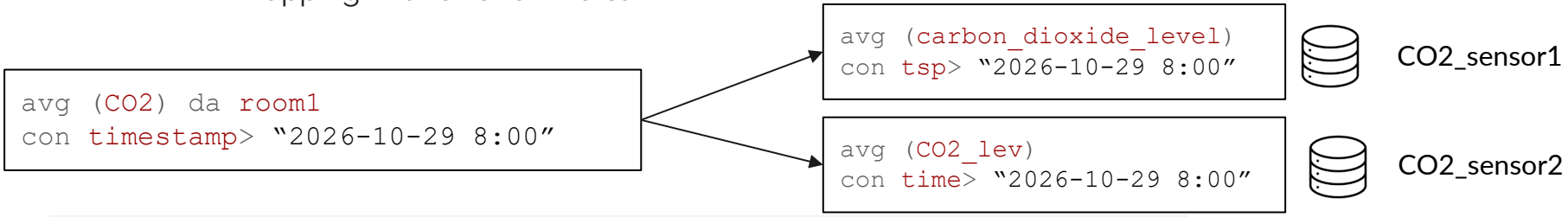
M. Arazzi, A. Nocera, E. Storti. "The SemIoE Ontology: A Semantic Model Solution for an IoE-based Industry." *IEEE Internet of Things Journal, Elsevier,* 2024.

Finanziato dall'Unione europea
NextGenerationEU

Industry 5.0

Applicazioni supportate:

- monitoring semantico di data stream, indipendente da eterogeneità legate a formato e schema dei dati
- query semantiche sui database dei sensori, con un approccio OBDA (Ontology-Based Data Access):
  - la query è espressa usando il vocabolario del Knowledge Graph
  - viene riscritta (*query rewriting*) per adattarsi alla specifica sorgente dati, usando i mapping in direzione inversa



M. Marconi Sciarroni, E. Storti. "A Context-Aware Knowledge Graph Platform for Stream Processing in Industrial IoT ." *in review, https://arxiv.org/abs/2602.19990.*

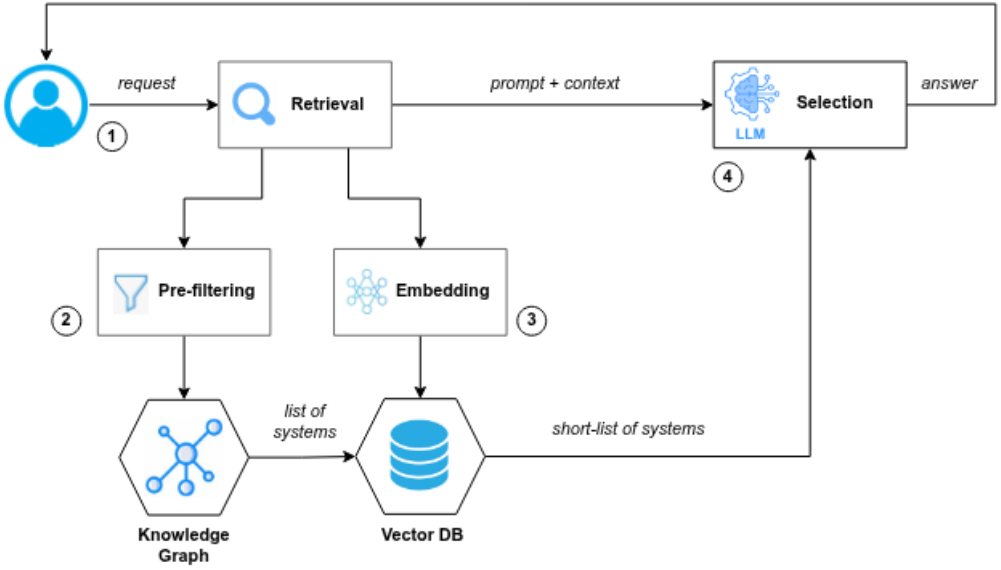
Finanziato dall'Unione europea
NextGenerationEU

Industry 5.0: KG + LLM

Framework basato su GraphRAG + Vector Store per estrazione di informazione context-aware:

1. La richiesta dell'utente è processata da un LLM

2. In base al contesto dell'utente (ruolo, permessi, localizzazione), il Knowledge Graph prefiltra i device accessibili



Arazzi, M., Marconi Sciarroni, M., Nocera, A., & Storti, E. RAG-IoE: IoT context-aware information retrieval with Large Language Models in Industry 5.0. *ACM Transactions on Internet of Things*, *6*(4), 1-31. 2025.

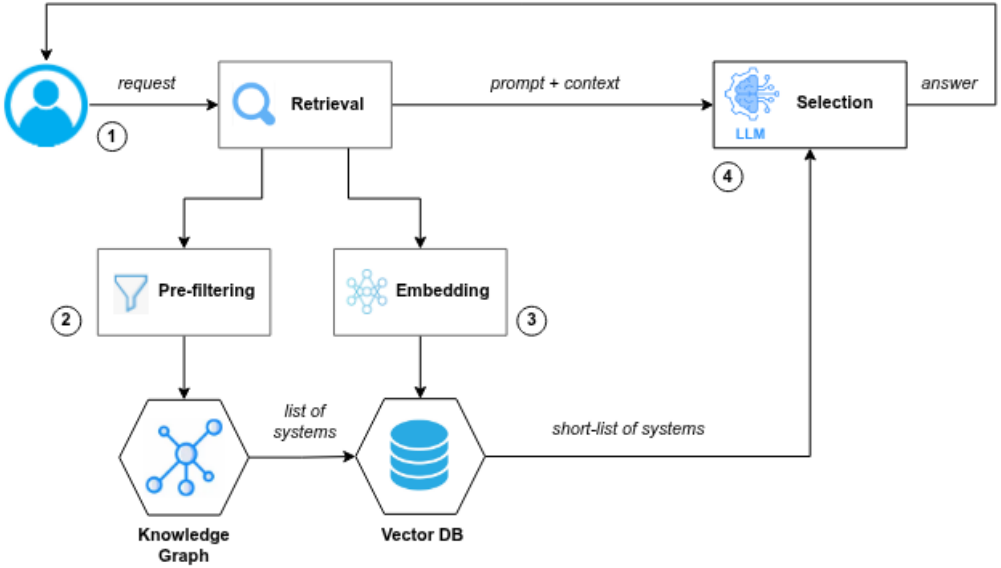
Finanziato dall'Unione europea
NextGenerationEU

Industry 5.0: KG + LLM

Framework basato su GraphRAG + Vector Store per estrazione di informazione context-aware:

3. Un'architettura RAG (Retrieval Augmented Generation) estrae l'insieme di device più rilevanti tra quelli prefiltrati e li ordina

4. Il LLM effettua la selezione e l'ordinamento finale

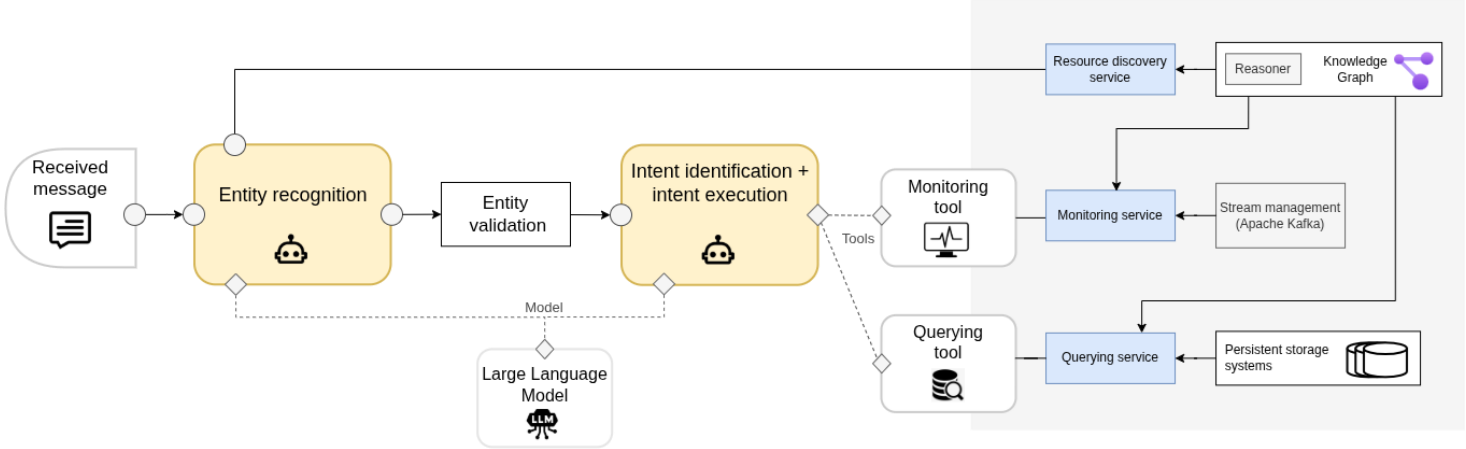


Arazzi, M., Marconi Sciarroni, M., Nocera, A., & Storti, E. RAG-IoE: IoT context-aware information retrieval with Large Language Models in Industry 5.0. *ACM Transactions on Internet of Things*, *6*(4), 1-31, 2025.

Finanziato dall'Unione europea
NextGenerationEU

Industry 5.0: Agenti AI context-aware

Agente AI per supportare l'accesso a tool della piattaforma (real-time streaming, query analitiche su DBMS) tramite linguaggio naturale e rispettando regole di accesso context-aware.

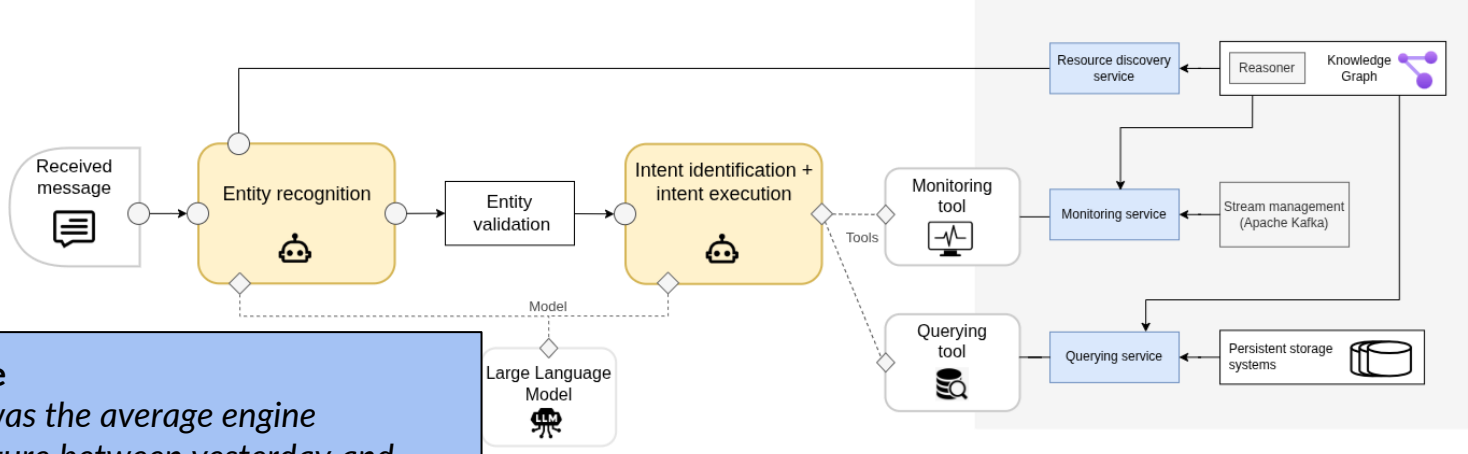


Marconi Sciarroni, M. & Storti, E. Agent-Based Service Orchestration for Context-Aware Data Access in Industrial IoT. In Proc of CAISE Workshops, LNCS, vol 586, 2026.

Finanziato dall'Unione europea
NextGenerationEU

Industry 5.0: Agenti AI context-aware

Agente AI per supportare l'accesso a tool della piattaforma (real-time streaming, query analitiche su DBMS) tramite linguaggio naturale e rispettando regole di accesso context-aware.



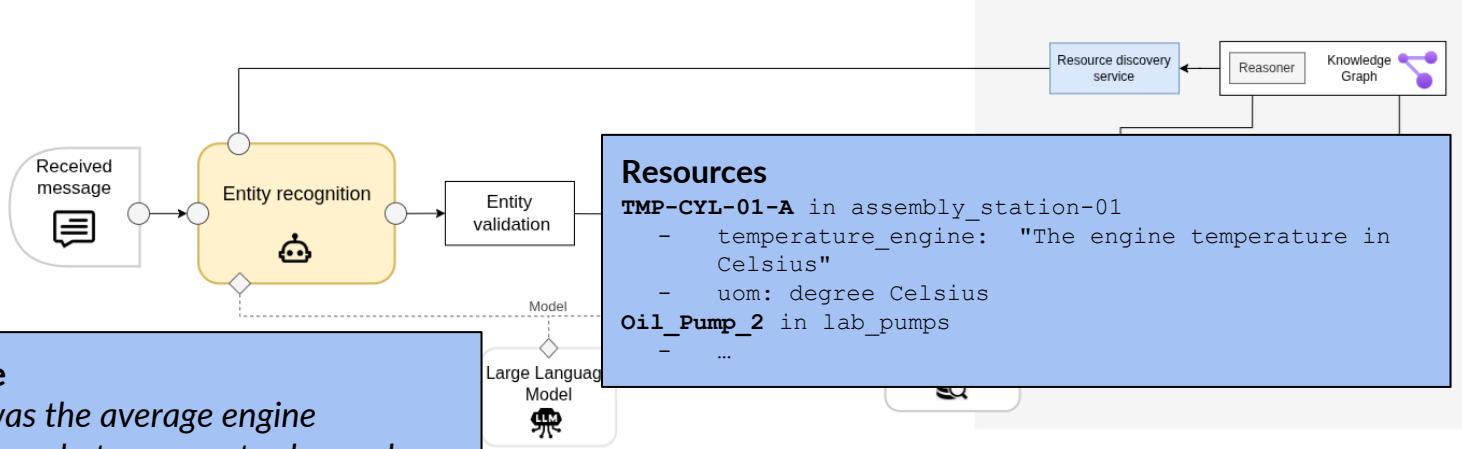
Message
<What was the average engine temperature between yesterday and today at noon for assembly station ?>>

d Service Orchestration for Context-Aware Data Access in Industrial
IoT. In Proc. of CAISE Workshops, LNCS, vol 586, 2026.

Finanziato dall'Unione europea
NextGenerationEU

Industry 5.0: Agenti AI context-aware

Agente AI per supportare l'accesso a tool della piattaforma (real-time streaming, query analitiche su DBMS) tramite linguaggio naturale e rispettando regole di accesso context-aware.

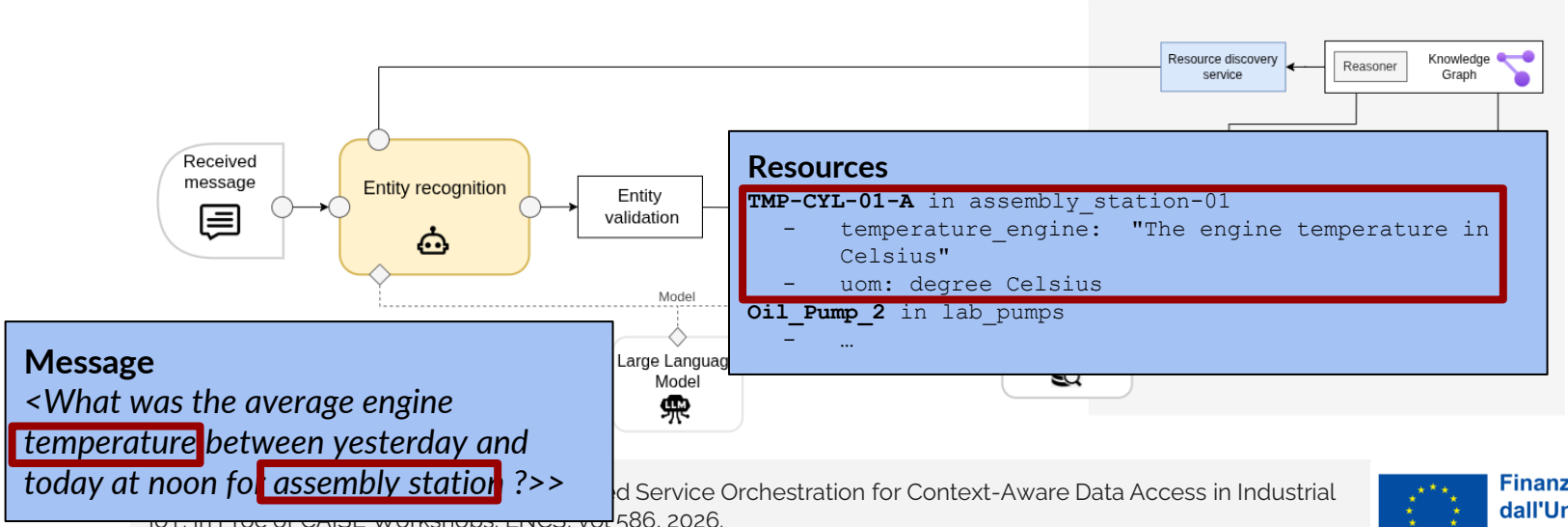


d Service Orchestration for Context-Aware Data Access in Industrial
IoT. In Proc. of CAISE Workshops, LNCS, vol 586, 2026.

Finanziato dall'Unione europea
NextGenerationEU

Industry 5.0: Agenti AI context-aware

Agente AI per supportare l'accesso a tool della piattaforma (real-time streaming, query analitiche su DBMS) tramite linguaggio naturale e rispettando regole di accesso context-aware.

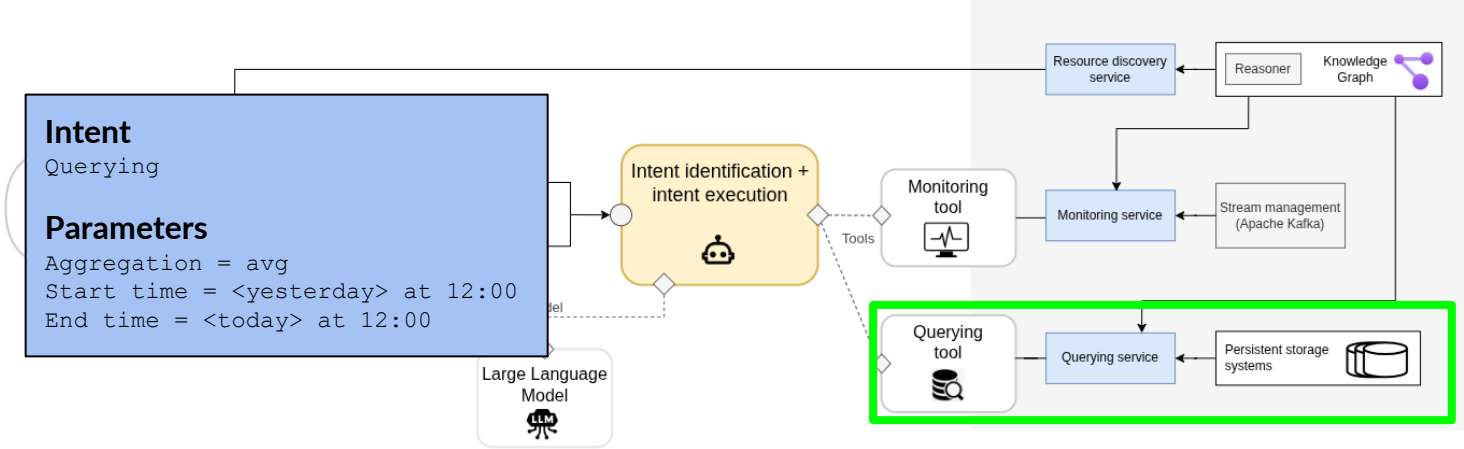


d Service Orchestration for Context-Aware Data Access in Industrial
IoT. In Proc. of CAISE Workshops, LNCS, vol 586, 2026.

Finanziato dall'Unione europea
NextGenerationEU

Industry 5.0: Agenti AI context-aware

Agente AI per supportare l'accesso a tool della piattaforma (real-time streaming, query analitiche su DBMS) tramite linguaggio naturale e rispettando regole di accesso context-aware.



Marconi Sciarroni, M. & Storti, E. Agent-Based Service Orchestration for Context-Aware Data Access in Industrial IoT. In Proc of CAISE Workshops, LNCS, vol 586, 2026.

Finanziato dall'Unione europea
NextGenerationEU

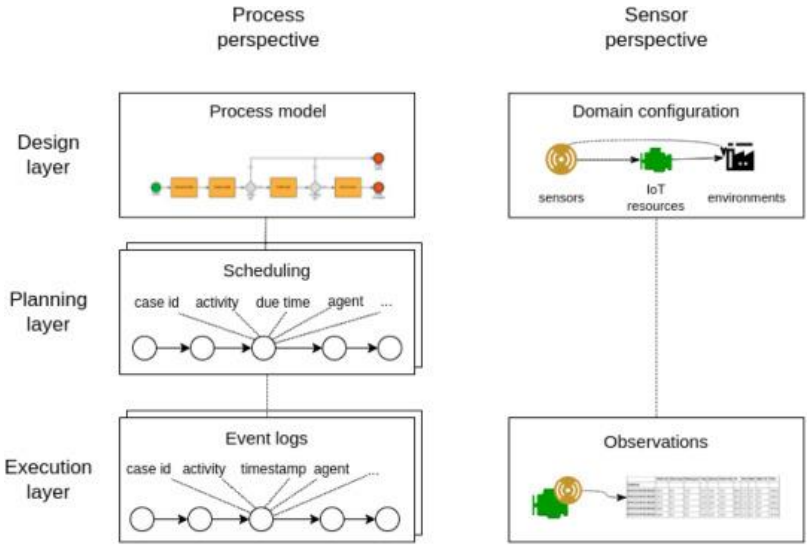
Verso “process-aware agentic AI”

Agentic AI per supportare l’analisi di processi

Rappresentazione:

- conoscenza di dominio (Knowledge Graph)
- sorgenti dati: database delle misurazioni dei sensori (DBMS), repository di documentazione tecnica (vector store)
- informazioni sui business process:
  - schema di processo
  - log dei processi eseguiti

Analisi: anomalie di processo, conformance checking, Root-Cause Analysis



Diamantini, C., Mircoli, A., Potena, D., & Storti, E. Process-aware IIoT Knowledge Graph: A semantic model for Industrial IoT integration and analytics. *Future Generation Computer Systems*, *139*, 224-238, 2023.

Emanuele Storti
e.storti@univpm.it