

INFN

National Institute for Nuclear Physics
Italy



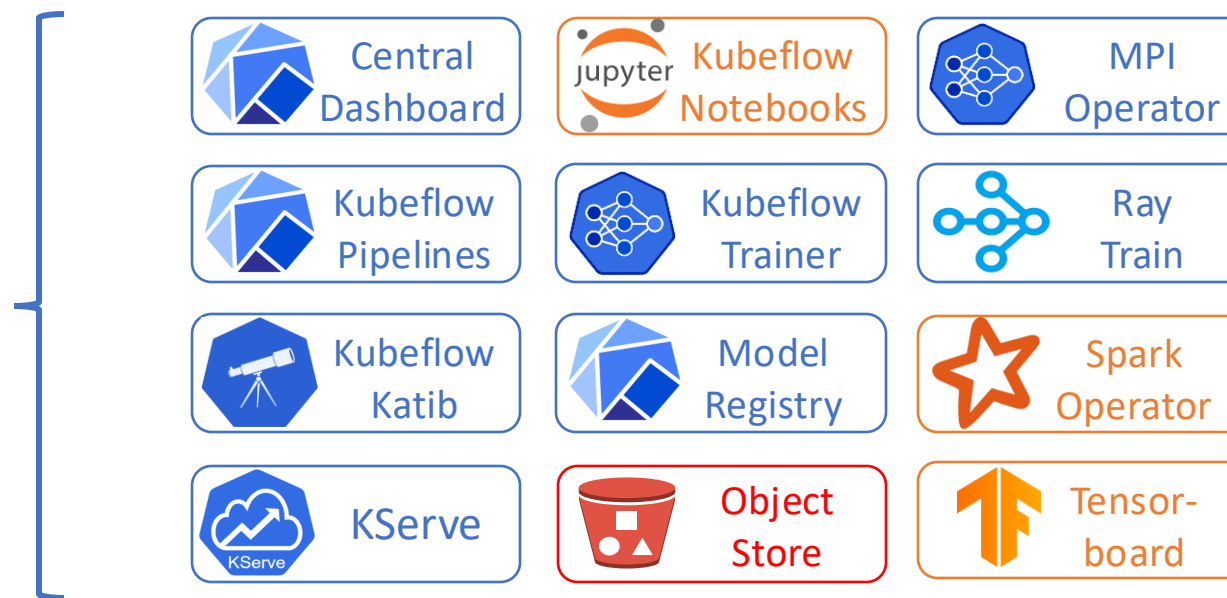
INFN Cloud Kubeflow as a Platform and use cases



Mauro Gattari
DSI/DataCloud
mgattari@infn.it

CCR Workshop
05/2025

Kubeflow

Open-source **Machine Learning** platform built on **Kubernetes** providing a set of tools to manage the whole lifecycle of an **ML solution**.



 Home Notebooks TensorBoards Volumes Katib Experiments KServe Endpoints Kotaemon (RAG) Model Registry Minio Pipelines

Manage Contributors

Multi-tenant

Kubeflow components

Log in to dex



Log in with Email



Log in with IAM for INFN Cloud

DEX
OIDC Provider



Central Dashboard

Side menu / Tenants / AuthN

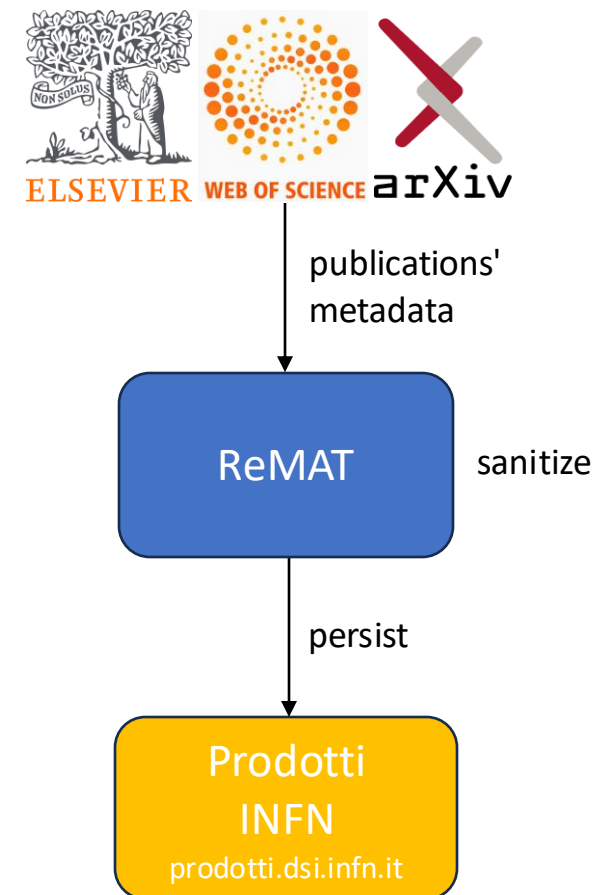
Use Case 1: ReMAT

Research Metadata Analysis Tool

In prodotti.dsi.infn.it we collect metadata from several sources.

Problem: metadata consistency, e.g.:

- aliases:
 - Rossi, Paolo Giovanni ✓
 - Rossi, PG ✓
 - Grossi, P ✗
- orcid:
 - 0000-0001-2345-6789 ✓
 - 0000-0001-XXXX-YYYY ✗
- affiliations:
 - INFN Frascati Natl Labs, I-00044 Frascati, Roma ✓
 - INFN Sez, Lab Nazl Frascati, Rome ✓
 - Univ Siena, Dipartimento Fis, Pisa, Italy ✗



ML Task

Classify Author's Affiliations

ML Task:

- Text Classification

Training dataset:

- ~6k **positive** samples
 - "INFN Frascati Natl Labs, I-00044 Frascati, Roma" -> LNF
 - "INFN Bari, Dept Phys, Bari, Italy" -> BA
- ~6k **negative** samples
 - "Univ Siena, Dipartimento Fis, Pisa, Italy" -> [Unknown]

Dataset augmentation:

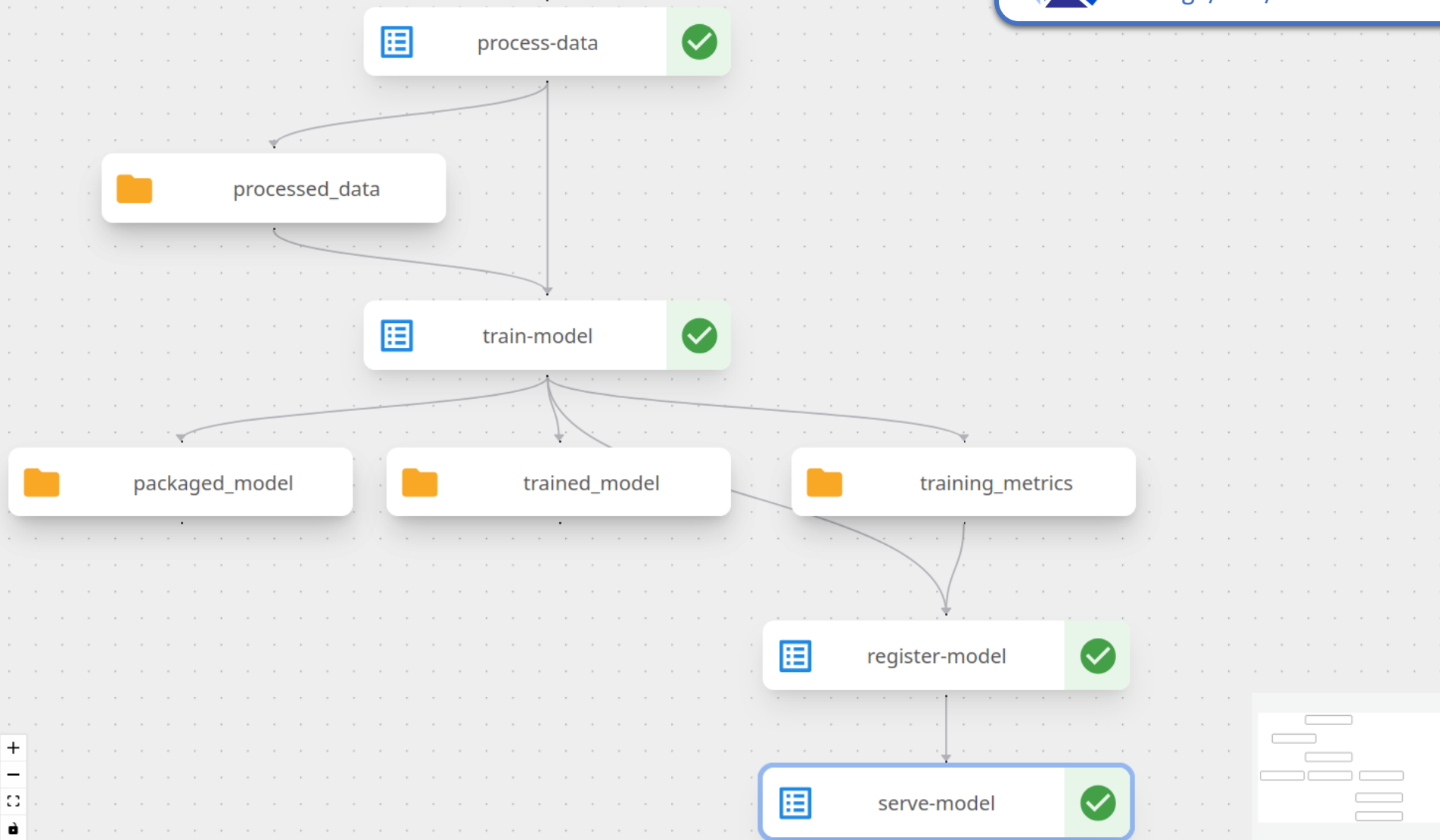
- ~400k synthetic samples by adding "smart" typos:
 - "1NFN Sez, Laab Nazl Frascati" -> LNF

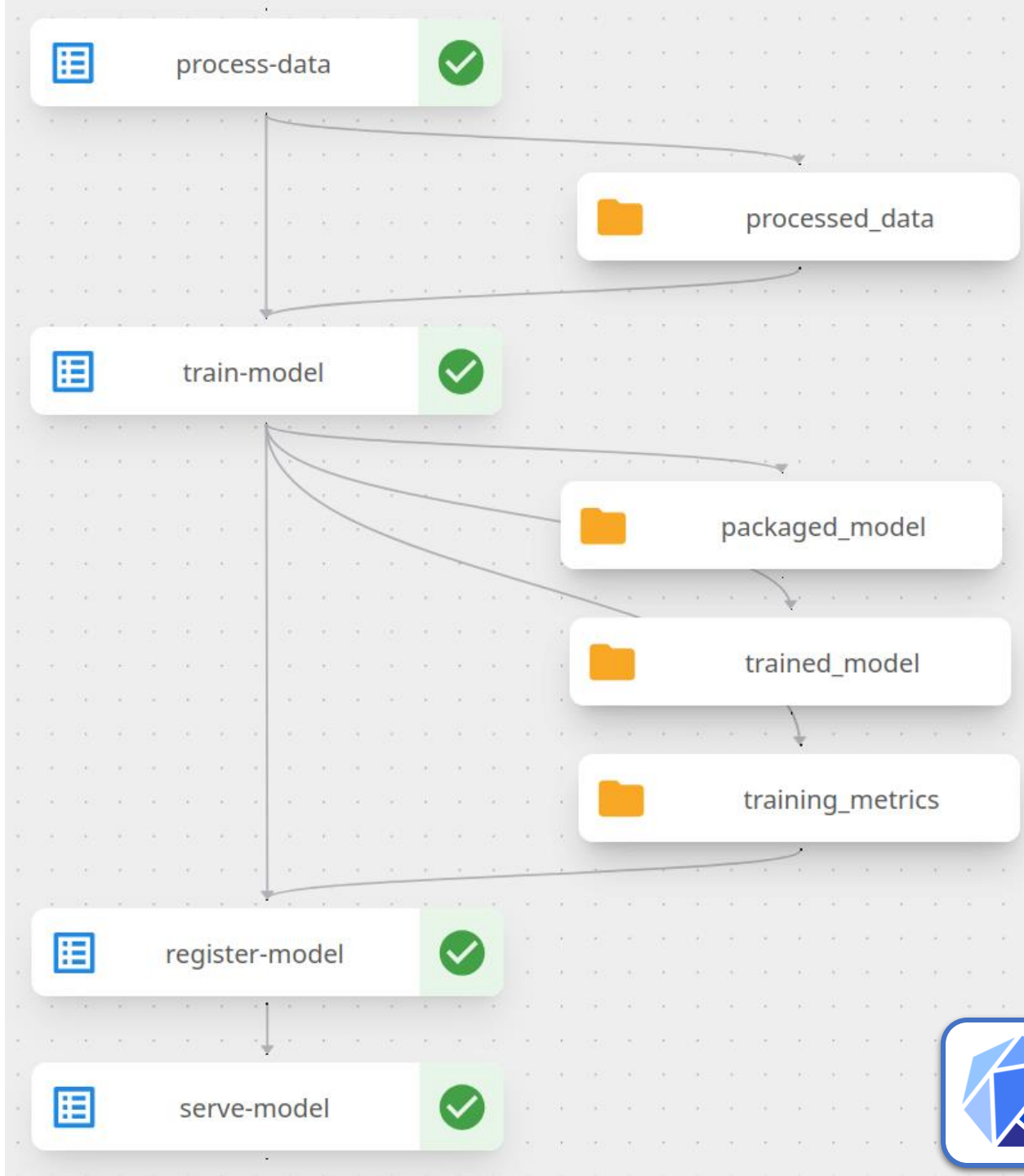
Training evaluation:

- 97% accuracy on test set

Kubeflow Pipelines

Design/Run/Schedule ML workflow





1.0.3

AffiliationsTC

Description

[EDIT](#)

Model trained to infer the INFN structure name given an author's affiliation string. The model was trained on ~12K samples with an accuracy on the test set of 97%.

Labels

[EDIT](#)

DSI

Properties

[+ ADD PROPERTY](#)

Key

Value

accuracy

0.97

⋮

author

mgattari@inf.n.it

⋮

model_format_name

pytorch

⋮

model_name

AffiliationsTC

⋮

model_registry_uri

model-registry://AffiliationsTC/1.0.3

⋮

model_storage_uri

s3://mlpipeline/v2/artifacts/affiliations-tc-pipeline/04e70904-54dc-4ad1-af6b-f310114ca17f/train-model/87d3fecc-3697-4bb4-bb08-1bbe61ce4cdb/packaged_model

⋮

model_version

1.0.3

⋮

[SHOW FEWER PROPERTIES](#)


Model Registry

Metadata central index

1.0.3

[ACTIONS](#)

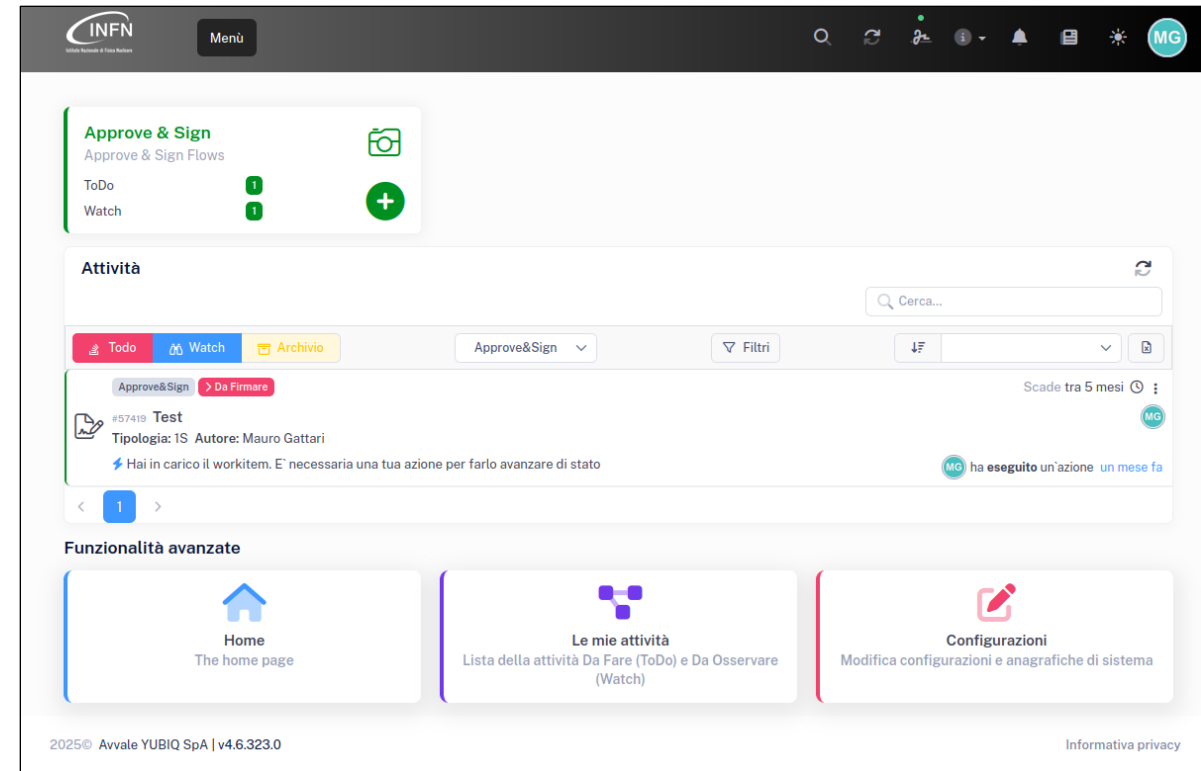
Use Case 2: ChatBot INFN LibroFirma

LibroFirma ChatBot

- AI assistant that answers user questions
- Knowledge Base:
 - ServiceDesk tickets
 - Transcription of "pillole formative" (<https://mediawall.infn.it/>)
- Fully-hosted: run on INFN Cloud resources

Generative AI

- Open-source LLMs (Large Language Models): provide "reasoning" capabilities
- Semantic Search: retrieve relevant information from the knowledge base to answer the question



ChatBot Language Models

- LLM (Text Generation):
 - **Alibaba/Qwen 2.5**
 - 72B parameters
 - ~60 tokens/sec (Nvidia A100 – thanks **AI_INFN**)
- Embeddings Model (Semantic Search):
 - **Snowflake/snowflake-arctic-embed-l-v2.0**
 - 568M parameters
- Reranker (Improve retrieval quality)
 - **BAAI/bge-reranker-v2-m3**
 - 568M parameters



AI Tools

- Kubeflow:
 - Design/implement/manage the AI solution
- Kotaemon:
 - Open-source application for Q&A with your documents

Kubeflow

Home

Notebooks

TensorBoards

Volumes

Katib Experiments

KServe Endpoints

Kotaemon (RAG)

Model Registry

Endpoints

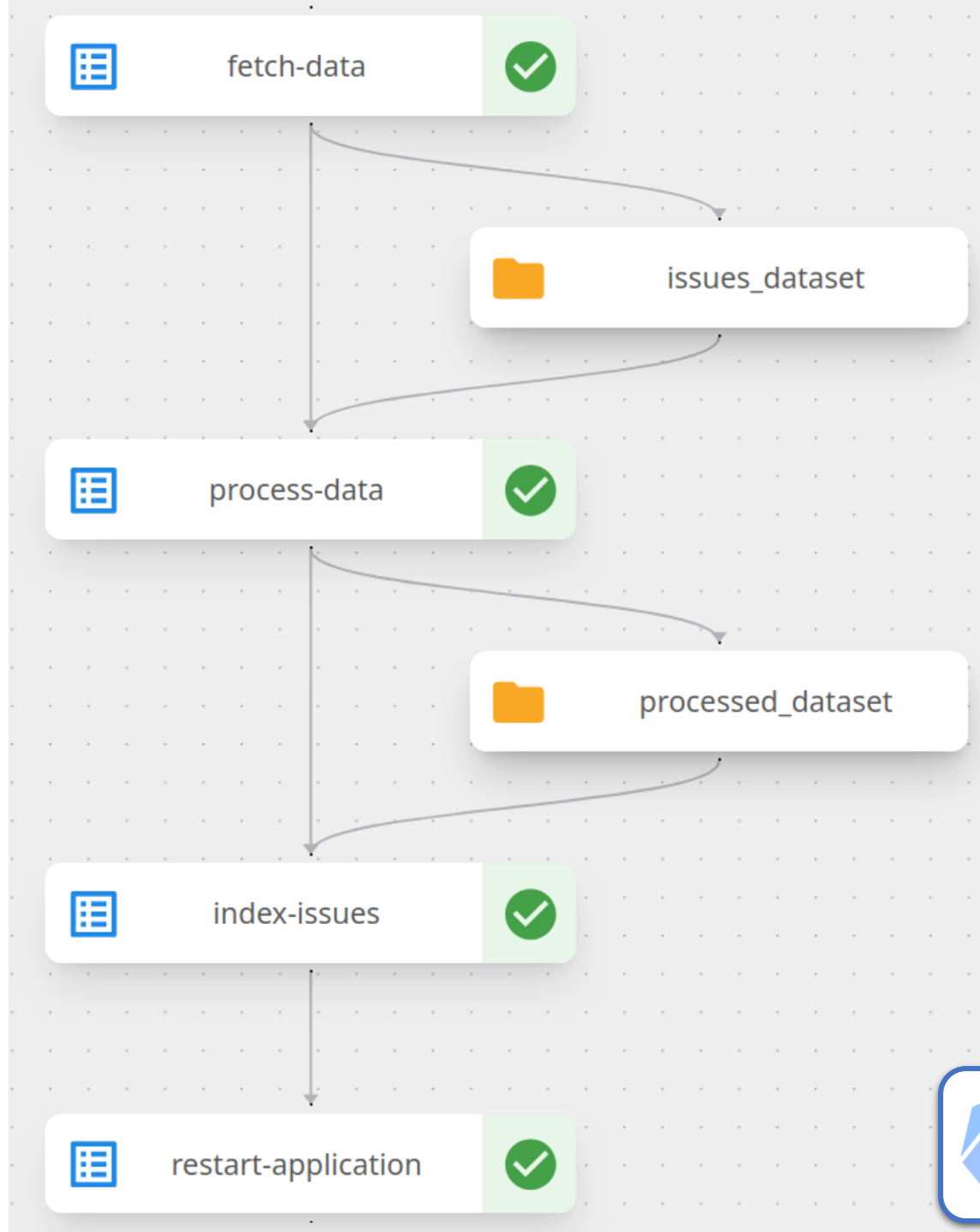
Filter Enter property name or value

Status	Name ↑	Created at	Predictor	Runtime	Protocol	Storage URI	
	chat-model	2 months ago	huggingface	vLLM	v1	s3://models/Alibaba/Qwen-2.5	
	reranker	5 days ago	huggingface	vLLM	v1	s3://models/BAAI/bge-reranker-v2-m3	
	text-embeddings	19 days ago	huggingface	vLLM	v1	s3://models/Snowflake/snowflake-arctic-embed-l-v2.0	

Items per page: 10 1 – 3 of 3

KServe
Inference Services

+ New Endpoint



mlaas (Owner)

ChatFilesResourcesSettingsHelp


Come faccio ad impostare la firma digitale?

Per impostare la firma digitale nel nuovo libro firma, segui questi passaggi:

- Accesso al menu dei certificati:**
 - Entra nella pagina principale e clicca sulle tue iniziali in alto a destra.
 - Nel menu che si apre, seleziona "I miei certificati" per procedere all'inserimento dei propri certificati di firma [1] .
- Aggiunta del certificato:**
 - Clicca sul pulsante "Aggiungi" per procedere [2] .
- Configurazione del certificato:**
 - Si aprirà una schermata chiamata "Nuovo Certificato".
 - Clicca sul menu a tendina di "Tipo" e seleziona "InfocertOTP" [2] .
 - In "Nome", inserisci un testo che ti aiuterà a identificare la tua firma digitale. Questo testo è un campo libera e ti sarà utile in futuro per riconoscere la tua firma [2] .
 - In "ID Utente", inserisci il testo "PAROSUNI", seguito dal simbolo "+" e poi dal tuo identificativo Infocert. Inserisci tutto senza spazi [2] .
 - Carica un'immagine che identificherà la tua firma digitale. Questa immagine sarà associata alla tua firma ogni volta che firmerai un documento [3] .
- Salvataggio del certificato:**
 - Clicca su "Salva Certificato" per completare l'operazione [3] .

Seguendo questi passaggi, avrai impostato la tua firma digitale nel nuovo libro firma Yubiq [1] [2] [3] .

Type a message, search the @web, or tag a file with @filename

 Kotaemon
LibroFirma ChatBot

Information panel

▼ Mindmap [Expand] [Export]

Come associare la firma digitale a Yubiq

Accesso alla pagina principale

Clicca sulle tue iniziali in alto a destra

Seleziona "I miei certificati"

Inserimento della firma elettronica semplice

Clicca su "Aggiungi"

Seleziona "Tipo" e scegli "Simple"

Inserisci un nome riconoscibile in "Nome"

Carica un'immagine di firma

Clicca su "Salva Certificato"

Inserimento della firma digitale

Clicca su "Aggiungi"

Seleziona "Tipo" e scegli "InfocertOTP"

Inserisci un nome riconoscibile in "Nome"

Inserisci l'ID Utente (PAROSUNI + identificativo Infocert)

Carica un'immagine di firma

Clicca su "Salva Certificato"

▼ NuovoLibroFirma**come**associare**la**firma**digitale**_mp4_720_.txt [score: 0.9]

► **Relevance score:** 0.9

Ciao, in questa pillola parleremo di come associare la propria firma elettronica, semplice e digitale, al nuovo libro firma Yubiq. Buon ascolto!

Per iniziare ad associare la firma, entra nella pagina principale. In alto a destra troverai le tue iniziali. Clicca su loro e si aprirà un menu. Ora seleziona i miei certificati per poter procedere all'inserimento dei propri certificati di firma. [2]

Poi clicca sul pulsante **Aggiungi** per procedere.

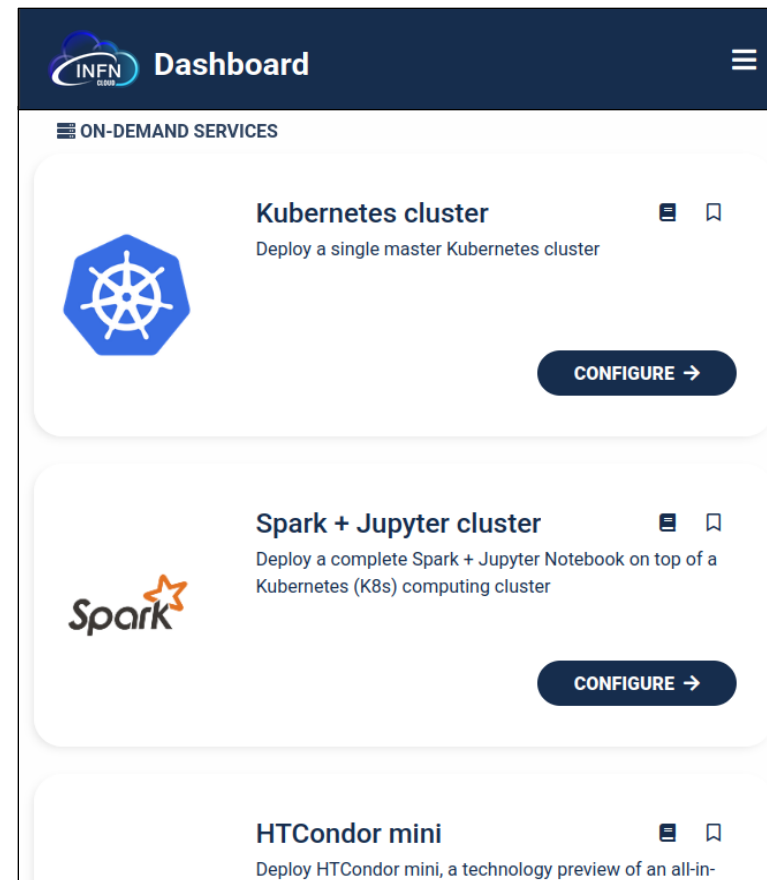
Si aprirà un menu chiamato Nuovo Certificato. Qui potrai impostare la tua firma elettronica semplice e la firma digitale. Iniziamo dalla firma semplice. Come prima cosa clicca sul menu a tendina di Tipo. Clicca sul menu la voce Simple. Ora, in Nome, [1] metti un testo libero che ti aiuti in futuro ad identificare la firma

What's next

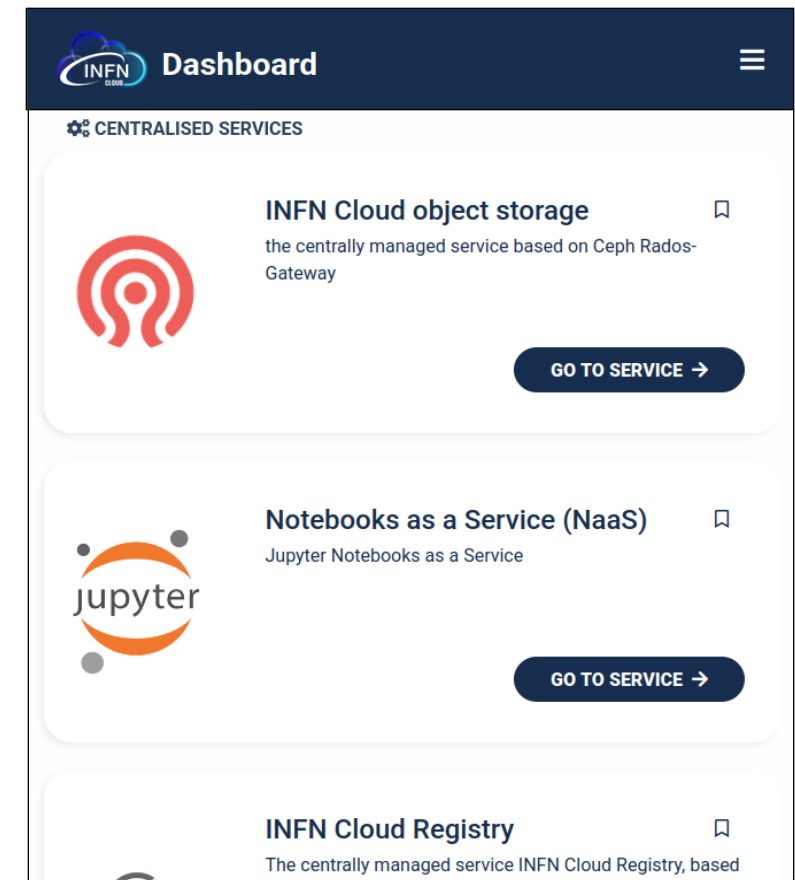
INFN Cloud Integration



2. On-Demand Service



3. Centralised Service



1. Manual install
 - hard, self-managed
2. On-Demand Service
 - easy, self-managed
3. Centralised Service
 - easy, centrally-managed
 - e.g. ml.cern.ch is a centralized service at CERN to run machine learning workloads

Thank you!

References:

- Kubeflow: www.kubeflow.org
- Kotaemon: github.com/Cinnamon/kotaemon

KaaP (Kubeflow as a Platform):

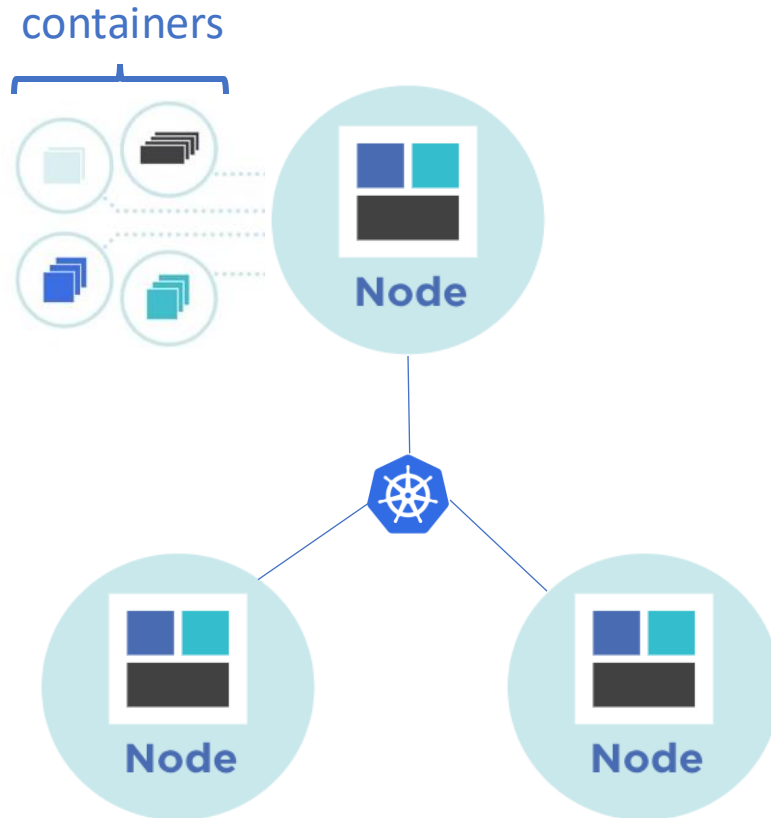
- Documentation: confluence.infn.it/Kubernetes Cluster with Kubeflow
- Source code: baltig.infn.it/kaap-manifests
- Manual Install (Ansible role): baltig.infn.it/ansible-role-kubeflow

ReMAT:

- Documentation: confluence.infn.it/Research Metadata Analysis Tool



Kubernetes

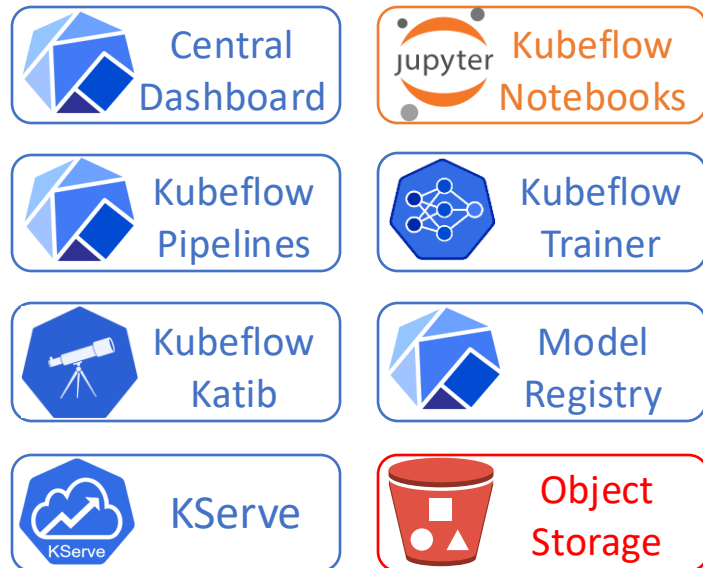


- **Open-source** technology for running **containerized** applications at scale.
- Providing features such as:
 - **Service Discovery**: enabling containers to find and communicate with each other.
 - **Load Balancing**: distributing traffic between containers.
 - **Scaling**: automatically scaling the number of running containers based on resources utilization.
 - **Self-Healing**: monitoring and restarting of failed containers.
 - ...

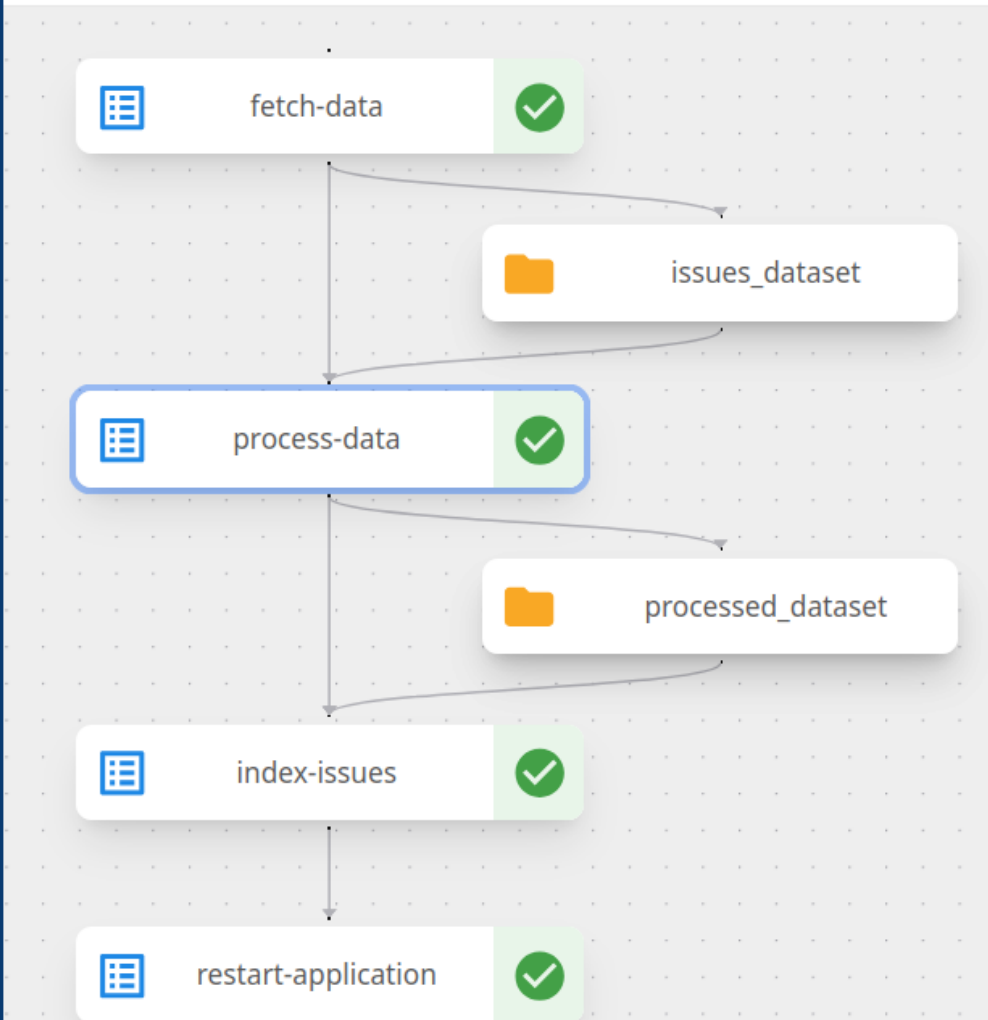
Kubeflow Ecosystem



Kubeflow Components



- Open-source **Machine Learning** platform built on **Kubernetes** providing a set of tools to manage the whole lifecycle of an **ML solution**.
- The **Kubeflow Ecosystem** of applications comprises the following:
 - **Central Dashboard**: web app for management of Kubeflow components.
 - **Notebooks**: web-based development environments.
 - **Pipelines**: orchestration tool to design and run ML workflows.
 - **Trainer**: distributed model training using TensorFlow, PyTorch, and other frameworks with support for GPU acceleration.
 - **Katib**: automatic hyper-parameter optimization.
 - **Model Registry**: central index to manage ML artifacts metadata.
 - **KServe**: tool for deploying ML models as scalable, reliable services.
 - **Object Store**: provides support for common storage technologies.



This step corresponds to execution "process-data".

Input Parameters

process_data_properties

1	{
2	"output_media_type": "text/plain"
3	}

Input Artifacts

issues_dataset

[minio://mlpipeline/v2/artifacts/libro-fir](#) [View All](#)
[ma-pipeline/629cf823-cef3-4121-865](#)
[1-4d1dec494d1a/fetch-data/a096d91b-](#)
[db86-4b70-9b85-0d6f45a8ef37/issues_](#)
[dataset](#)

Output Parameters

nr_of_issues	200
nr_of_messages	1036
result	"success"

Output Artifacts

processed_dataset

[minio://mlpipeline/v2/artifacts/libro-fir](#) [View All](#)
[ma-pipeline/629cf823-cef3-4121-865](#)
[1-4d1dec494d1a/process-data/68c221](#)
[25-863a-46f1-811b-00540725b705/pro](#)
[cessed_dataset](#)

User settings

Retrieval settings

Reasoning settings

Save & Close

Max context length (LLM)

32000

Language model

Qwen2.5 (default)

QA Prompt (contains {context}, {question}, {lang})

Sei un assistente AI che aiuta gli utenti a risolvere i loro problemi relativi al software 'Libro Firma' di gestione della firma digitale.

Rispondi in modo chiaro e dettagliato alla domanda dell'utente, usando il seguente contesto che riporta informazioni estratte dal manuale utente del 'Libro Firma' ed una serie di tickets aperti da vari utenti e gestiti dall'assistenza.

Se non sai la risposta, rispondi che non hai sufficienti informazioni per rispondere, non tentare di formulare una risposta.

Se l'utente formula una domanda relativa ad argomenti non pertinenti al software del 'Libro Firma', rispondi che non sei progettato per fare conversazioni generiche.

Rispondi sempre in italiano.

{context}

Domanda dell'utente: {question}

Risposta utile:

Number of interactions to include

5



Kotaemon
Reasoning settings

Conversations

HybridRAG: Advanced Information Retrieval

Shared

File Collection

Search All

Search In File(s)

GraphRAG Collection

Search All

Search In File(s)

HybridRAG- Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction.pdf

Quick Upload

Drop File Here

- or -

Click to Upload

What is HybridRAG

HybridRAG is an advanced approach that combines the strengths of two methodologies: VectorRAG and GraphRAG, to enhance the quality, accuracy, and comprehensiveness of responses in large language models (LLMs). It is specifically designed to retrieve relevant information from external documents, particularly unstructured financial texts, to provide more accurate answers.

Key Features of HybridRAG:

1. Integration of VectorRAG and GraphRAG:
- VectorRAG:

Utilizes vector-based retrieval to find relevant information, especially useful for generating responses to abstractive questions where answers are not explicitly mentioned in the data.
- GraphRAG:

Leverages Knowledge Graphs (KGs) to enhance information extraction capabilities, particularly effective for extractive questions where answers can be directly pulled from the data.
2. Enhanced Information Extraction:
- By combining the broad, similarity-based retrieval of VectorRAG with the structured, relationship-rich retrieval of GraphRAG, HybridRAG can efficiently extract and interpret complex information from financial documents and other sources.
3. Improved Response Generation:

Chat Input

Send

Regen

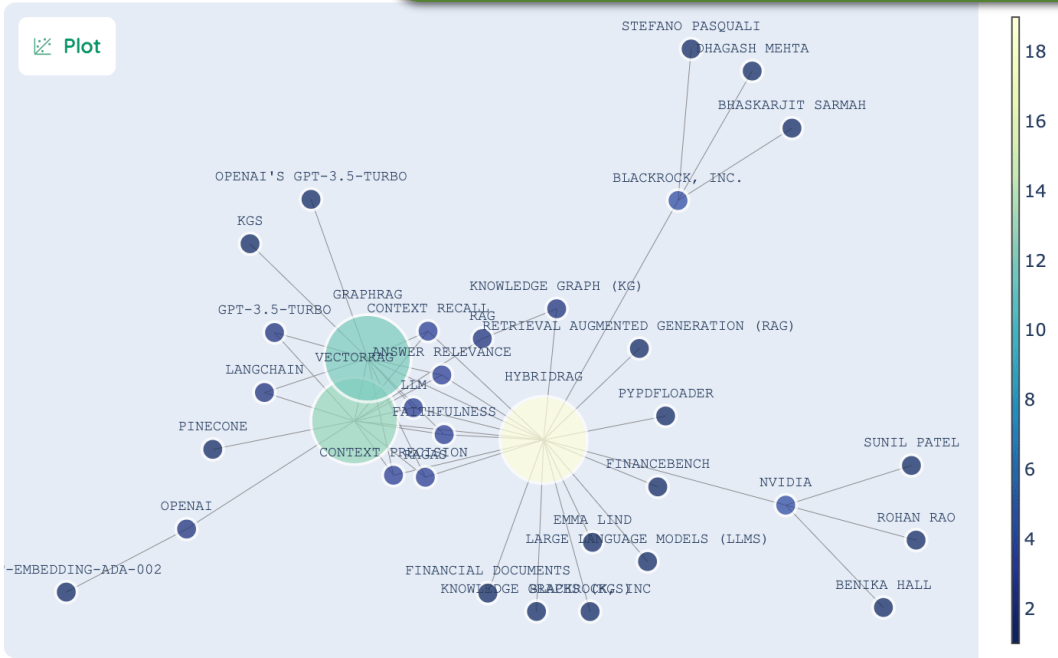
Chat settings



Kotaemon

Full Picture

Information panel



▼ Table from **Entities** [score: 1.0]

► **Relevance score:** 1.0

entity	description
HYBRIDRAG	HybridRAG is a novel and integrated approach that combines the strengths of VectorRAG and GraphRAG methodologies to improve the quality, accuracy, and comprehensiveness of responses in large language models (LLMs). It is specifically designed to retrieve relevant information from external documents for queries to LLMs, aiming to provide more accurate answers by leveraging the strengths of both RAGs. This system is particularly focused on the efficient extraction and interpretation of complex information from unstructured financial texts. By integrating Knowledge Graphs and Vector Retrieval, HybridRAG enhances information extraction and response generation, making it a promising approach for balancing high-quality answers with comprehensive context retrieval in information extraction tasks. This technique is described as an innovative solution for integrating knowledge graphs and vector retrieval augmented generation, showcasing its capability in efficiently extracting information from financial documents and other sources. Overall, HybridRAG represents