ML-AAS

Machine Learning

As a Service

ML-AAS

AI Platform
providing
Services in the Cloud

- **Train/fine-tune** machine learning models at scale
- **Host/share** datasets and trained models in the cloud
- Serve models to make **predictions** about new data
- Manage models and versions through a **public INFN catalog**

AI Platform

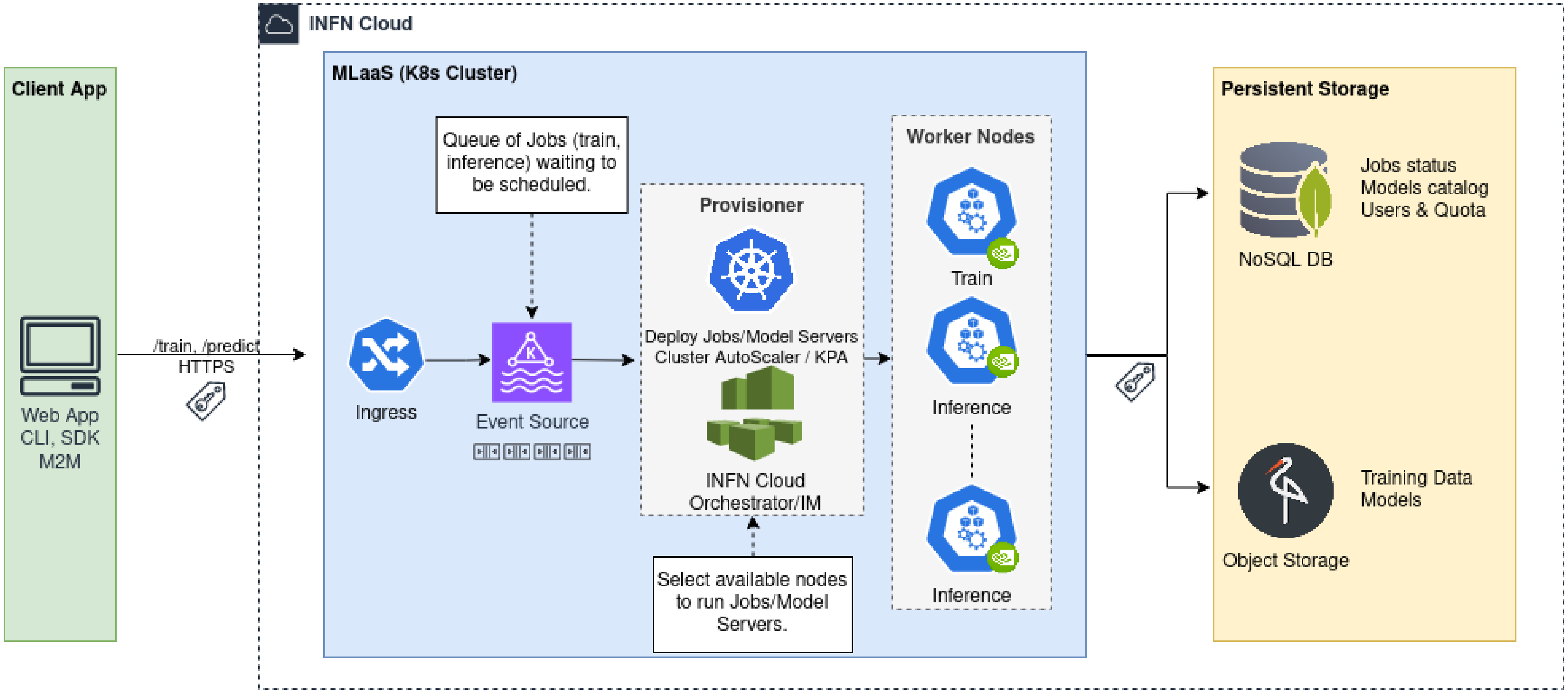
Building Blocks - Technologies

- **Computing resources:** CPUs, RAM, GPUs + Networking – **INFN Cloud**
- **Container Orchestrator**: automate deployment, scaling, and management of workloads on physical/virtual nodes – **Kubernetes**
- **Event Source**: decouple jobs submissions from their execution - **Kafka**
- **Provisioner:**
  - **Train**: run distributed training jobs, hyperparameters tuning – **Kubeflow Training Operator, Kueue, Katlib**
  - **Inference**: models serving – **KServe, Knative + KPA, Batching, Deployment Strategies, Inference Pipelines**
  - **Cluster Scaling** - **K8s AutoScaler, INFN Cloud Orchestrator/IM**
- **NoSQL DB**: keep jobs status; maintain public catalogs - **Mongo DB**
- **Object Storage**: host data and models – **S3, MinIO, Longhorn**
- **Client Apps/Tools**: Web App, CLI, SDKs, etc. to accelerate integration with the platform

on top of
INFN Cloud

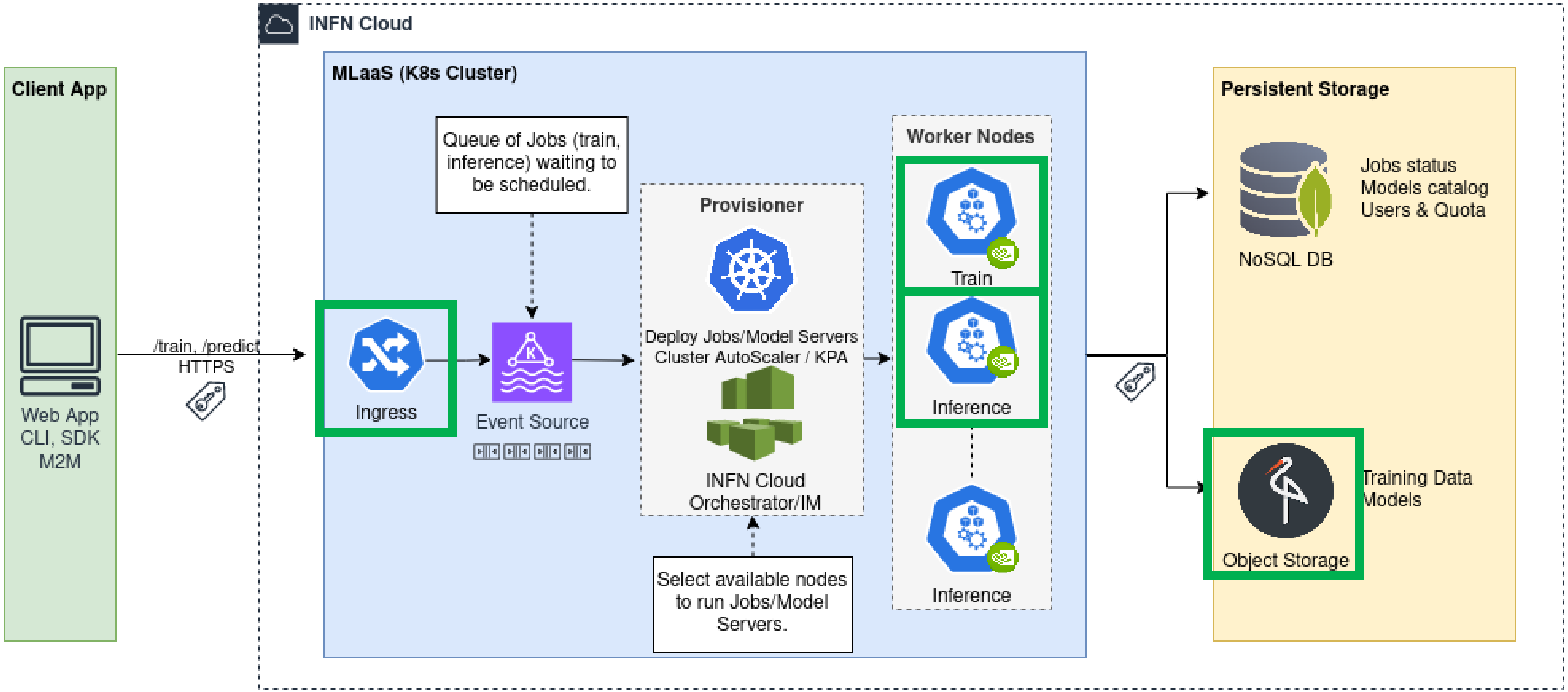
AI Platform

High Level Architecture – TO BE



AI Platform

High Level Architecture – AS IS

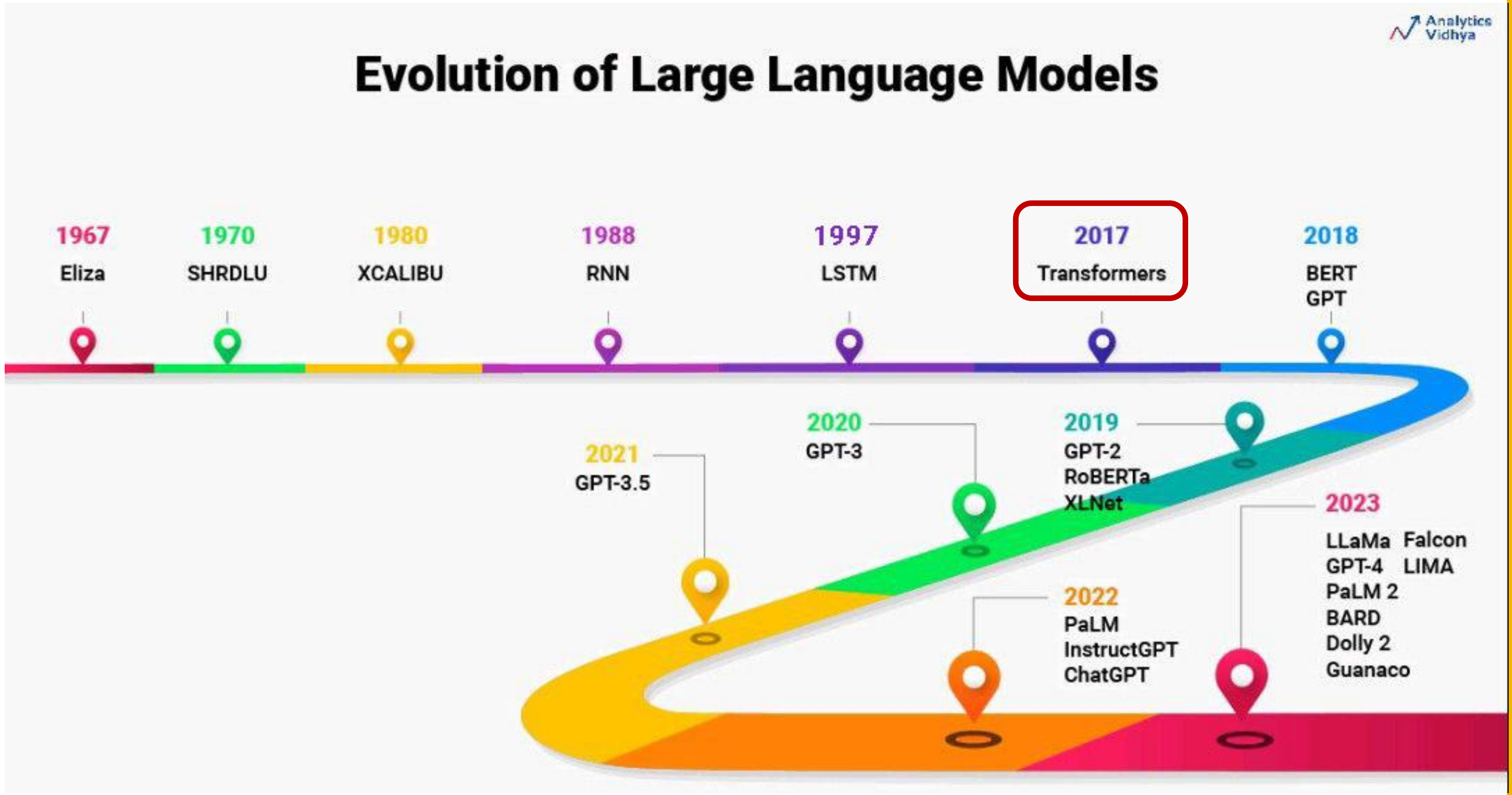


ML-AAS RESTful API

- **POST** `/api/v1/ml/cv/train/model` Train a CV Model
- **POST** `/api/v1/ml/cv/predict` Get predictions with a pretrained CV Model
- **POST** `/api/v1/ml/nlp/train/model` Train a NLP Model
- **POST** `/api/v1/ml/nlp/train/tokenizer` Train a Tokenizer Model
- **POST** `/api/v1/ml/nlp/predict` Get predictions with a pretrained NLP Model
- **POST** `/api/v1/ml/nlp/search` Semantic search with a pretrained NLP Model
- **POST** `/api/v1/ml/nlp/vector/search` Search vectors in a Vector Store index
- **POST** `/api/v1/ml/nlp/vector/index` Embed documents and store resulting vectors in a Vector Store index
- **GET** `/api/v1/ml/utilities/s3/object` List S3 objects

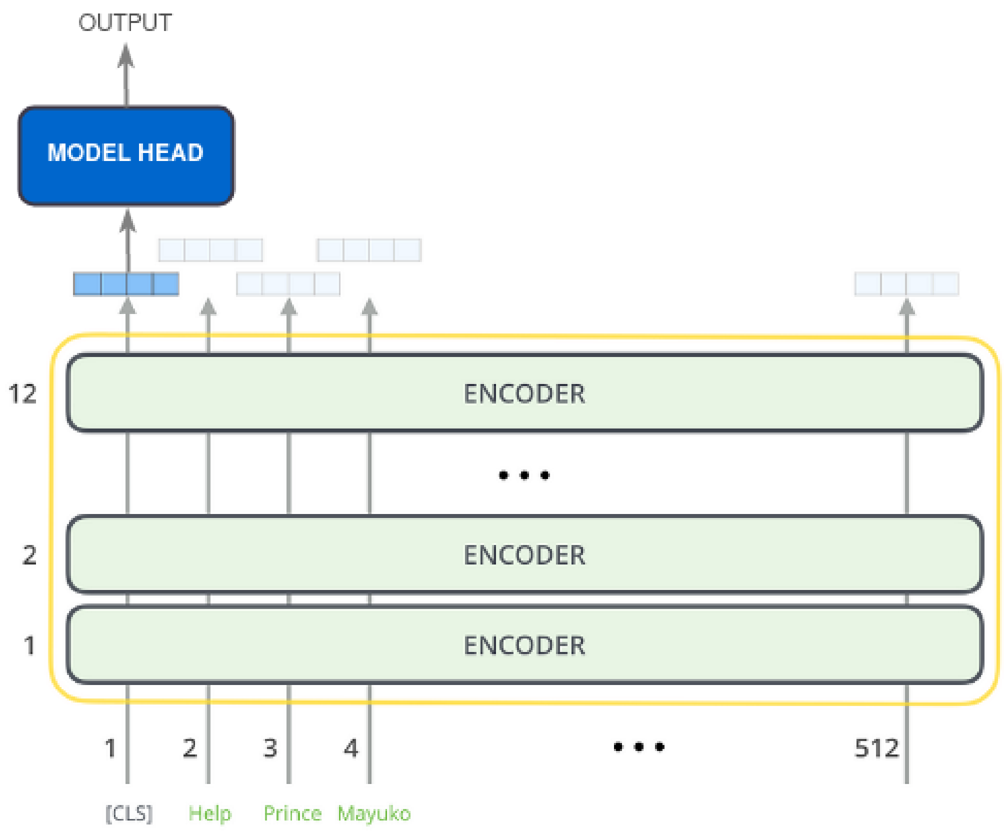
NLP

- NLP is a field of linguistics and machine learning focused on understanding the human language.
- A key event in the history of Language Models is the introduction of the Transformers architecture in June 2017 (Google Brain team).
- Transformer models have proven to be very powerful in solving NLP tasks.



Transformer Architecture

- Stack of **encoding layers** (now we are not interested in the decoding stack)
- Each layer captures a different level of **linguistic information**, from surface features to deep semantic features in the higher layers
- Each layer outputs an **embedding vector** for each input sentence token – a dense vector that represents the contextual understanding of the token/sentence by the Transformer model
- Last layer's output is given to a custom **Model Head** designed for a specific task, e.g. Text Classification



Text Classification

Use case:

- predict INFN structure name from author's affiliation string

Training:

- **model**: distilbert
- **dataset**: labeled author's affiliation strings:
  - ~6k positive samples
  - ~6k negative samples
  - dataset augmented to ~400k samples by adding "smart" typos
- **training**:
  - 6 hours (Nvidia Tesla T4)
  - 95% accuracy

Author's affiliation	INFN Structure
Catania Univ, Ist Nazl Fis Nucl, Lab Naz Sud, Catania, Italy	CT
Bari INFN, Via E Orabona 4, I-70125 Bari, Italy	BA
Dell INFN Frascati, Lab Nazl, Frascati, Italy	LNF
INAF IASF Milano, I-20133 Milan, Italy	[NoSite]

Training Dataset

Text Classification

Let's try on ML-AAS

```
{
  tokenizer: {
    path: path/to/tokenizer,
    storage_type: s3 | hugging_face_hub | local
  },
  model: {
    path: path/to/model,
    storage_type: s3 | hugging_face_hub | local,
    objective: text-classification
  },
  predict_input: {
    input_text: [
      "Frascati Natl Lab INFN LNF, Natl Inst Nucl Phys, Italy"
    ]
  }
}
```

Masked Language Modeling

Use case:

- fine-tune the Language Model to understand the semantics of sentences about physics

Training:

- **model**: bert, longformer
- **dataset**:
  - ~60k publication abstracts
  - 10% masked tokens

Sentence	Masked Sentence
The determination of the spin-parity properties of the discovered Higgs Boson is one of the main goals of the ongoing analyses at LHC. This note describes the experimental technique used by the ATLAS collaboration to test different spin-parity hypotheses [...]	The determination of the spin-parity **[Masked]** of the discovered Higgs Boson is one of the main **[Masked]** of the ongoing **[Masked]** at LHC. This note describes the experimental technique used by the ATLAS collaboration to test **[Masked]** spin-parity hypotheses [...]

Training Dataset

MLM Models

Model	Nr. Hidden Layers	Nr. Parameters	Training Time (10 epochs – 16M tokens)	Accuracy	Max Input Length
BERT	12	109M	8:24:03	71%	512
BERT Large	24	335M	1 day, 2:20:31	75%	512
LONGFormer	12	148M	1 day, 1:43:29	74%	4096
LONGFormer Large	24	434M	3 days, 7:26:04	78%	4096

MLM

Let's try on ML-AAS

```
{
  tokenizer: {
    path: path/to/tokenizer,
    storage_type: s3 | hugging_face_hub | local
  },
  model: {
    path: path/to/model,
    storage_type: s3 | hugging_face_hub | local,
    objective: masked-lm
  },
  predict_input: {
    input_text: [
      "Particles have corresponding antiparticles with the same mass
   but with [MASK] electric charges. Thus, the positron, which is a
   positively [MASK] [MASK], is the antiparticle of the negatively
   charged electron."
    ]
  }
}
```

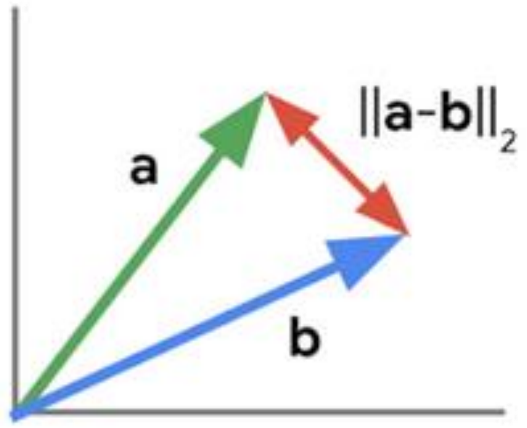
Semantic Search

Use Case:

- find the INFN project that "best" matches a publication abstract

Vector Store Search:

- **embeddings**: MLM-tuned Language Model – dense vectors that capture the semantics of a sentence;
- **dataset**: ~20k publication abstracts;
- **vectore store**: FAISS - Facebook AI Similarity Search - efficient storage and searching for embeddings;
- similarity search with score (L2 distance – lower is better).



L2 distance

Vector Search

Let's try on ML-AAS

```
{
  tokenizer: ...,
  model: {
    path: path/to/model,
    storage_type: s3 | hugging_face_hub | local,
    objective: no-objective
  },
  dataset: {
    path: path/to/vector/store/to/load,
    format: vector_store,
    storage_type: s3 | local
  },
  search_input: {
    input_text: [
      "Angular correlations between charged trigger and ..."
    ]
  }
}
```

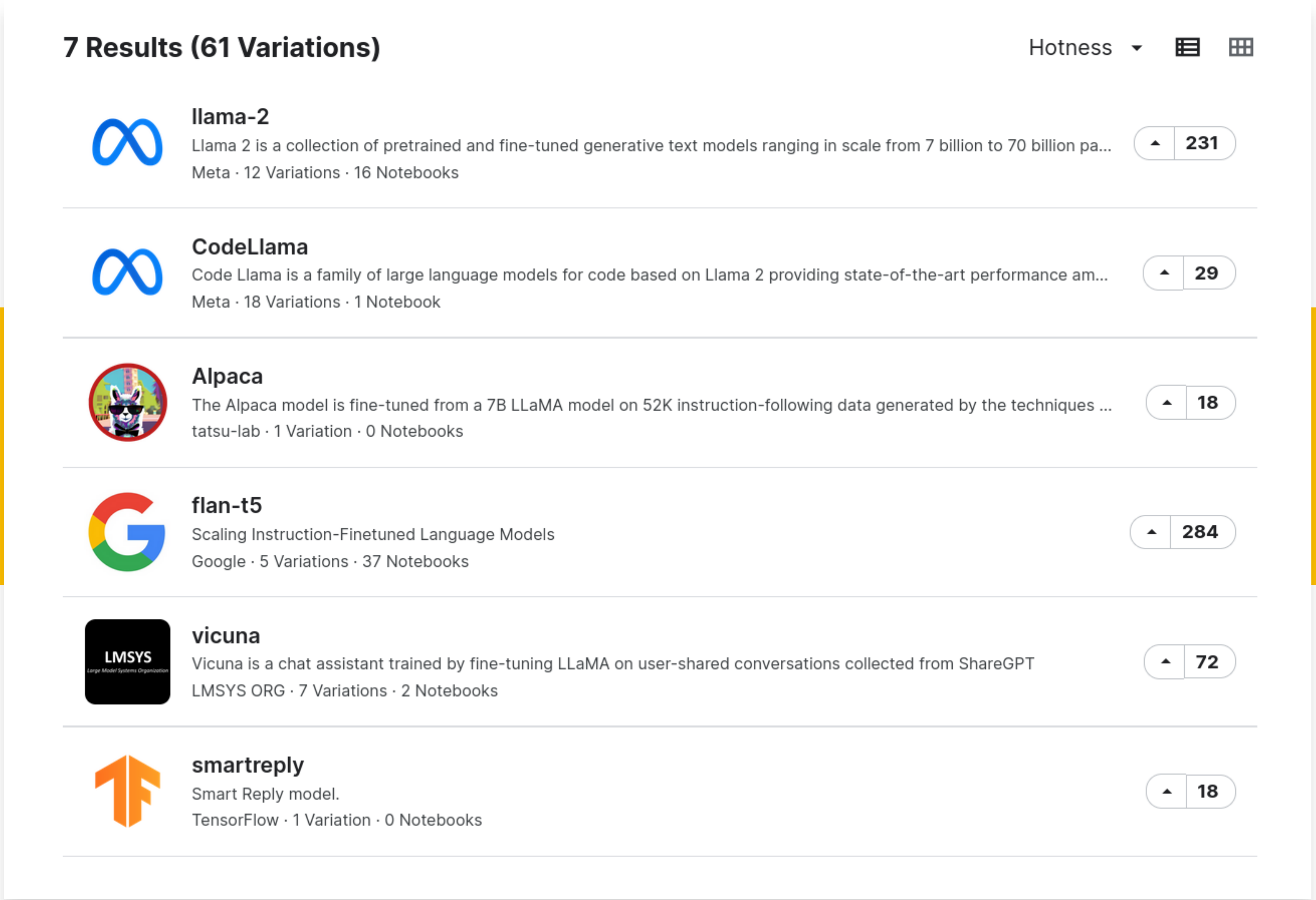
Training
Input

```
{
  tokenizer: ...,
  model: {
    path: path/to/model/to/load,
    storage_type: s3| hugging_face_hub  | local,
    objective: causal-lm  | masked-lm | next-sentence-prediction  | text-classification  | ...
  },
  model_train: {
    epochs: 10,
    batch_size: 32,
    optimizer: {
      name: "AdamW",
      init_lr: 2e-5,
      num_warm_steps: 1000,
      weight_decay_rate:  0.01
    }
  },
  dataset: {
    path: path/to/dataset/to/load,
    storage_type: s3 | local,
    train_test_split: 0.1
  },
  model_save: {
    path: path/to/model/to/save,
    storage_type: s3 | local
  }
}
```

Training

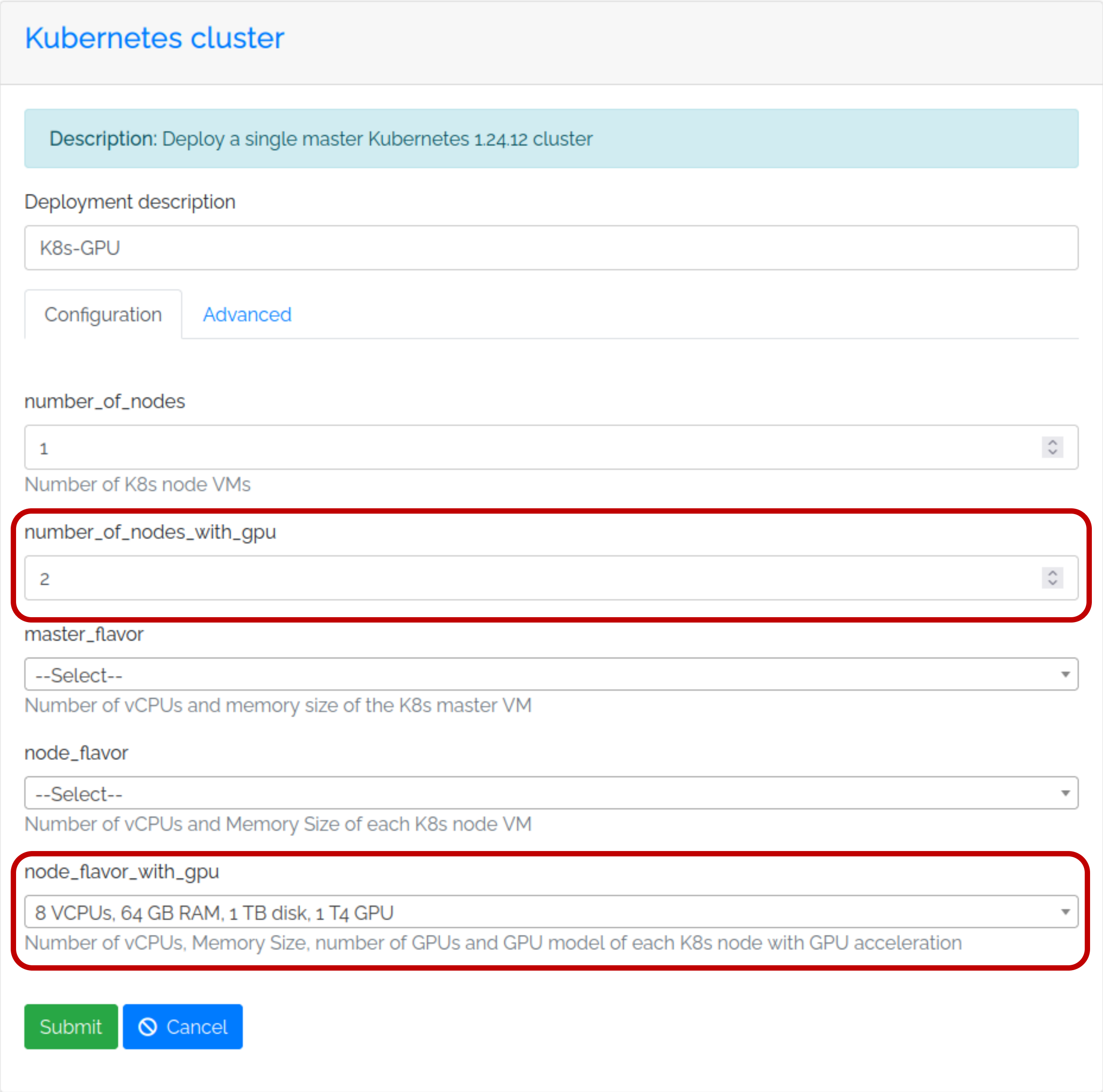
Output


```
{
   task_id: "0751bf4c-95f7-4463-bc47-dd901561e1df",
   task_status: succeeded,
   stats: {
      submitted: "2023-09-13T06:30:48",
      elapsed: "1 day, 2:20:31",
      ...
   },
   history: {
      loss: [1.67, 1.46, ..., 1.17],
      accuracy: [0.67, 0.70, ..., 0.75]
      ...
   },
   evaluation: {
      loss: 1.16,
      accuracy: 0.76,
      ...
   },
   dataset: {
      samples_train: 51227,
      samples_test: 6325,
      tokens: 13756616,
      ...
   }
}
```



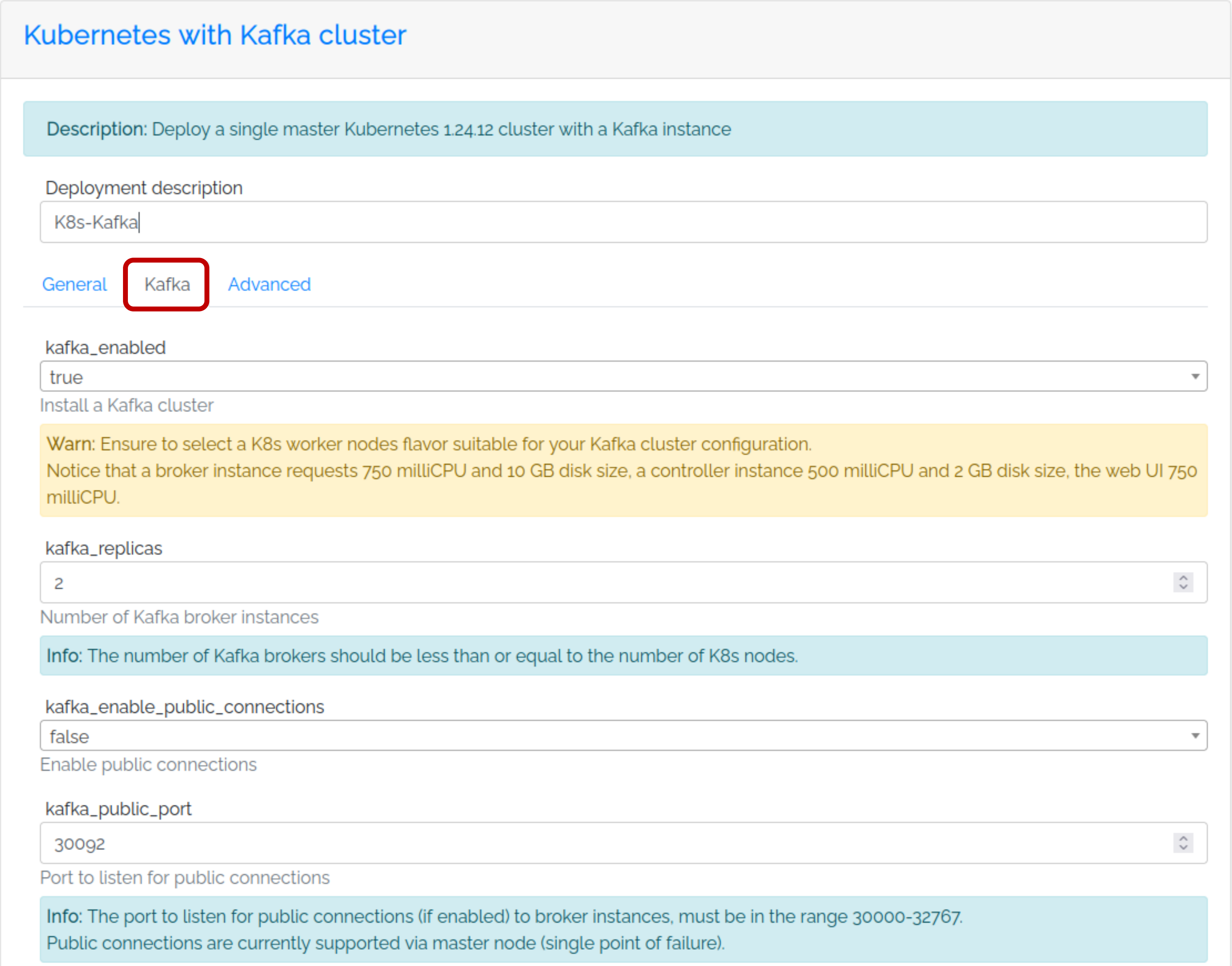
Web App

Models Catalog



INFN Cloud

K8s GPU Support



INFN Cloud

K8s Kafka Support

INFN Cloud
ML-AAS

INFN Cloud Managed PaaS
Service

Deploy a private K8s cluster
with ML-AAS

TODO

Architecture

- Finalize platform architecture and technologies
- Develop INFN Cloud connectors, e.g. K8s Auto-Scaler

NLP

- Implement other NLP objectives, e.g. NER, Question-Answering, etc.
- Integrate non Transformer-based models

NLP – Use Cases for **INFN Research Products catalog**

- Consider smaller/larger language models
- Collect more data for training

ML

- Add ML use cases to support INFN core research, e.g. ML for HEP (High Energy Physics)

...

Thank You

Democratizing AI

- Share data, algorithms, computing resources, and knowledge
- Provide tools to automate and accelerate the lifecycle of an AI project
- Reduce time and cost of AI development, increase productivity
- Promote collaboration and openness, foster creativity
- Promote widespread adoption of AI

Use Cases

Training on data extracted from INFN publications.

- **Text Classification**

  Predict INFN structure name from author's affiliation string:

  - **"CNAF, Ist Nazl Fis Nucl, Bologna, Italy"** -> **CNAF**
  - **"CSDC, Sez INFN Firenze, Florence, Italy"** -> **FI**

- **MLM (Masked Language Modeling)**

  Predict missing tokens in sentences about physics:

  - **"The determination of the spin-[Masked] properties of the discovered Higgs Boson..."** -> **parity**

- **Semantic Search**

  Find the INFN project that "best" matches a publication abstract:

  - **"The black hole images obtained with the Event Horizon..."** -> **CSN4/Teongrav**
  - **"We investigate the density distributions acquired by a..."** -> **CSN2/Fish**

Vector Index

```
{
  tokenizer: ...,
  model: {
    path: path/to/model,
    storage_type: s3 | hugging_face_hub  | local,
    objective: no-objective
  },
  dataset: {
    path: path/to/dataset/to/load,
    storage_type: s3 | local,
    loader_kwargs: {
      page_content_column: column-name
    }
  },
  vector_store_save: {
    path: path/to/vector/store/to/save,
    format: vector_store,
    storage_type: s3 | local
  }
}
```

Python Dependencies

Web Server:

- FastAPI, Uvicorn, KServe

ML:

- Tensorflow, Keras, Transformers, Datasets, Evaluate, Scikit-learn

Vector Index/Search:

- LangChain, FAISS, Doctran

Other:

- Pydantic, Numpy, Pandas, Boto3, typo, clean-text, graphviz, ...