2. odprti dan za gospodarstvo

SLAIF
Slovenska tovarna UI
Slovenian AI Factory

# Predstavitev primerov uporabe storitev SLAIF v podjetjih

***dr. Daniel Vladušič,***

***XLAB d.o.o.***

10. April 2026


# XLAB d.o.o

Get IT done.

Microsoft Partner
Silver Application Development

Red Hat
Advanced
Business Partner

DELOVE PODJETNIŠKE ZVEZDE 2021

ISO/IEC 27001 I-059

EU programme:
Horizon Europe

# IT rešitve / v partnerstvu z naročnikom

**XLAB**

Razvoj Programske Opreme

**2001**

Ustanovljeni v Tehnološkem Parku Ljubljana

**60+**

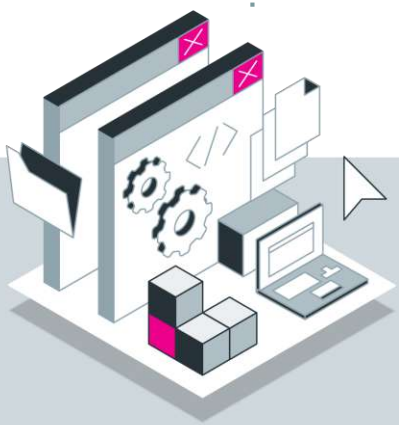
zaposlenih

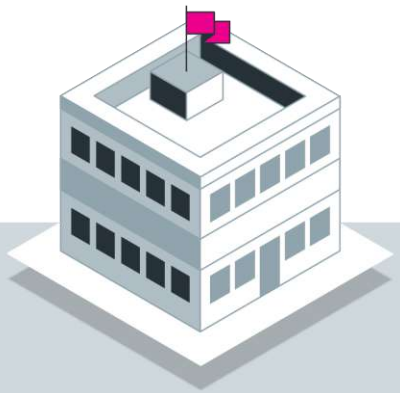
**Top 50***

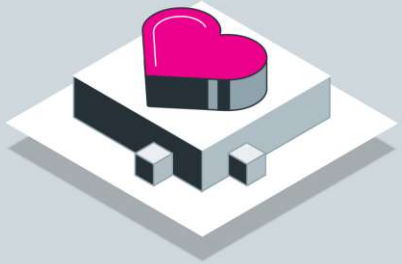
EU raziskave

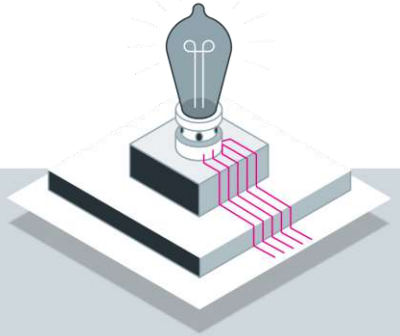
**Get IT done.**

Slogan









# XLAB Produkti & Raziskave

XLAB Steampunk

- IT Automation with Ansible
  - Ansible Playbook Scanning Tool

Gaea+

3D GIS & Visualization

XLAB MEDICAL

3D Medical Imaging

## Raziskovalni oddelek

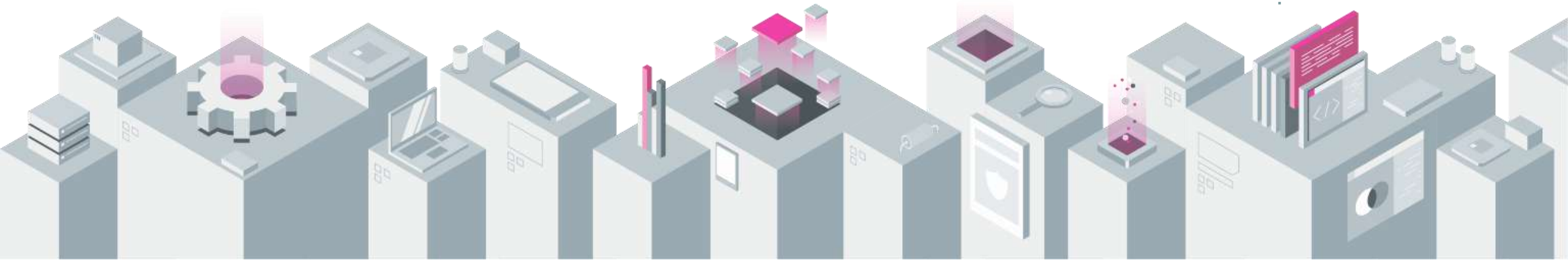
Eden največjih v Sloveniji

# Ekspertna področja

IT Automation &
Management

AI &
Machine Learning

Application
Modernization



# Naslovljeni izziv v sklopu SLAIF

- Smo dolgoletni uporabniki klasičnih HPC storitev.
- Sodelovali v EU projektih – HPC za SME (Fortissimo)
- **Razumemo, da za uporabo HPC potrebuješ eksperta (system access, Torque, Slurm, etc.)**
- SLAIF obljublja enostavnejši pristop:
  - **običajni razvojnik & LLM/AI ekspert.**

# Naslovljeni izziv v sklopu SLAIF

Imamo relevanten primer uporabe (LLM fine-tuning):

Ansible Code Suggestion

**Q: Kakšne so prednosti SLAIF vs. HPC?**

# Spotter & LLM

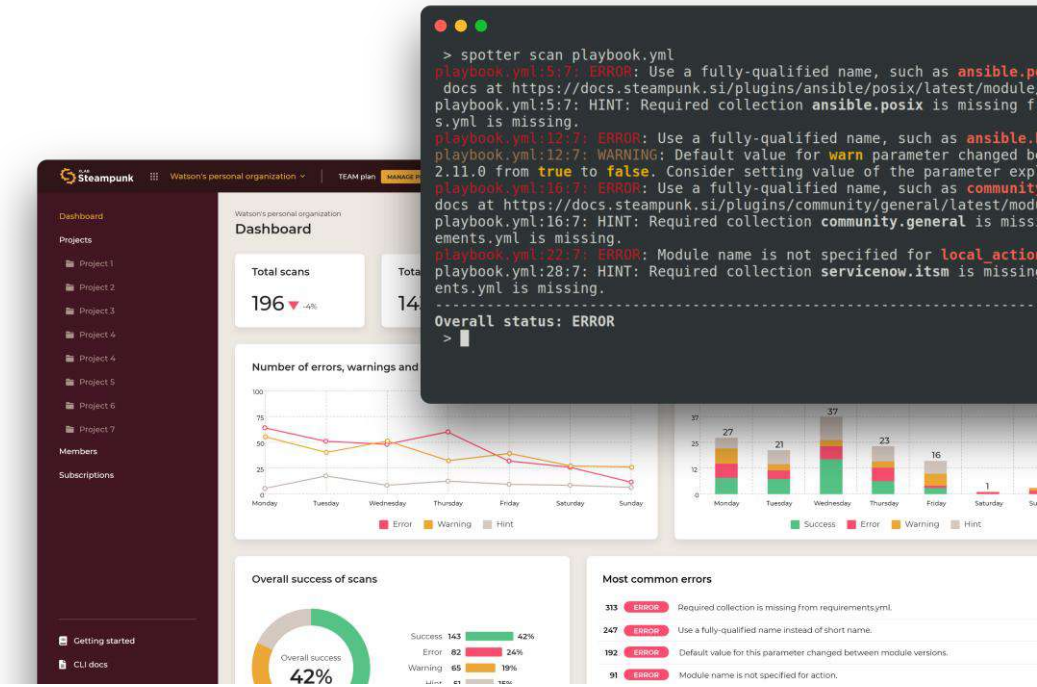
Steampunk Spotter

**Trustable Automation**

Ansible Playbook Platform that scans, analyzes, enhances, and provides insights for your playbooks.

XLAB

- **Produkt:**
  - Spotter - analizira in izboljšuje infrastrukturno kodo (infrastructure as a code - IaC). **Ansible kot osnovni jezik za IaC.**
- **Kako:**
  - Skenira playbooke (skripte), zazna morebitne težave, svetuje, kako jih izboljšati.
- **Zakaj:**
  - Odprava tveganj v osnovnih gradnikih porazdeljenih sistemov.



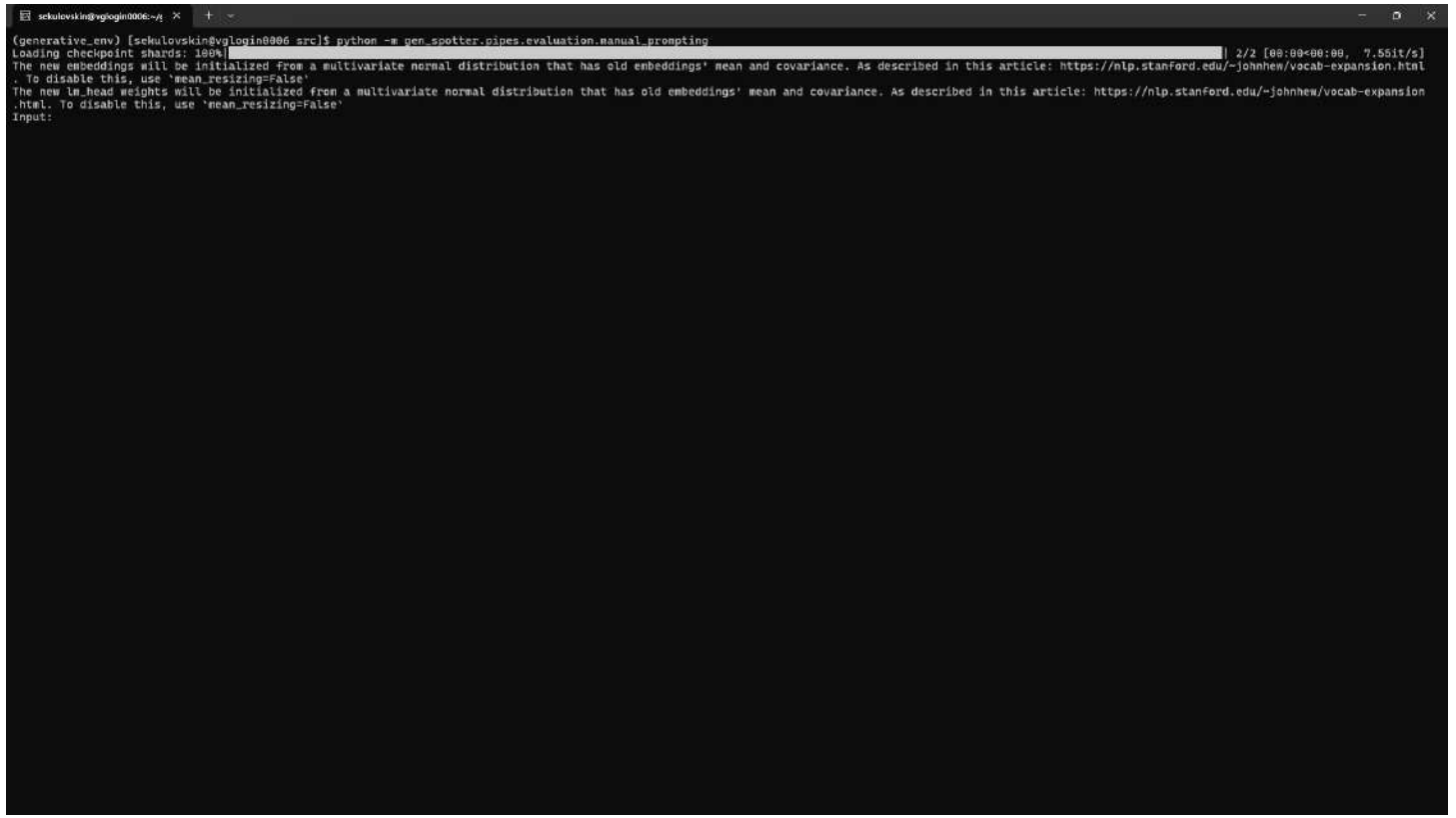

Red Hat · MindPoint Group · Island Networks · Messer Gases for Life · OceanBridge

# Zakaj LLM za Ansible skripte?

- **Ekspertov za IaC skoraj ni.**
- **IaC je ključna za pravilno (virtualno) infrastrukturo**

- Manjko učnih podatkov
- Manjko “zmogljivosti“
- On-prem namestitev

- LLM je pomoč ekspertu – ne nadomestilo **(HITL)**



# Naslovljeni izziv v sklopu SLAIF


**Kakšne so prednosti SLAIF vs. HPC?**

# Izkušnje

| PODROČJE | UGOTOVITVE |
| --- | --- |
| **Podatki** | JSONL smo brez težav naložili preko Storage API. |
| **Uporaba podatkov (virtualka / StorageAPI)** | Nalaganje/prenašanje posameznih/več datotek je potekalo brez težav. |
| **Priprava modela** | Za delo bi koristila instanca s prednameščenim CUDA toolkitom. |

# Izkušnje

| PODROČJE | UGOTOVITVE | POMEN |
|---|---|---|
| **Temeljna razlika glede na HPC** | Namesto zaklenjenega HPC sistema z omejenimi pravicami smo imeli root dostop do virtualne instance v OpenStack okolju. | Več fleksibilnosti pri nameščanju knjižnic, prilagoditvah sistema in pripravi lastnega razvojnega okolja. |
| **Shranjevanje podatkov** | Storage API je dobro zasnovan; deluje kot ločen podatkovni sloj in hkrati kot de-facto varnostna kopija. | Bolj varen in pregleden workflow, posebej za podatke in checkpoint-e. |
| **Delo z datotekami** | Možen je tudi izpis oziroma prenos samo dela datoteke namesto cele datoteke. | Učinkovitejša EDA in manj nepotrebnega prenosa podatkov. |

# Naša opažanja / prilagoditve

**Podpora za podatkovne tipe:** Dodeljene grafične kartice so starejše generacije in ne podpirajo podatkovnega tipa bf16, zato smo uporabili fp16.

**Namestitev CUDA okolja (CUDA Toolkit):** Virtualna instanca ni imela prednameščenega CUDA Toolkita, za kombinacijo nameščene različice CUDA gonilnika in različice Ubuntu pa ni bilo na voljo združljivega paketa.

**Prilagoditve kode:** Namesto postavljanja nove instance smo prilagodili kodo obstoječemu okolju. Brez CUDA Toolkita ni bilo mogoče namestiti knjižnic flash_attn_2 in DeepSpeed, zato smo namesto njiju uporabili SDPA za mehanizem pozornosti in DDP za porazdeljevanje učenja.

**Trenutni workflow:** Podatkovne datoteke se najprej prenesejo iz sistema za upravljanje podatkov, nato se izvede učenje modela, po zaključku pa se imeniki s kontrolnimi točkami naložijo nazaj v sistem za upravljanje podatkov.

# Priporočila

| | |
|---|---|
| **Uporabniška navodila** | Eksplicitna navodila za uporabnika v trenutni fazi SLAIF. |
| **Priprava okolja** | Uporabnik naj si vnaprej pripravi ponovljivo okolje (uv & lock file, Docker image, ...). |
| **CUDA instanca** | Na voljo naj bo instanca s CUDA Toolkitom in znanimi specifikacijami (različica OS, CUDA, ...). |
| **Gola instanca** | Na voljo naj bo tudi gradnja gole instance, kjer vešč uporabnik vse namesti sam. |

# Ocena

| | | |
|---|---|---|
| **Splošna ocena** | V trenutni fazi **SLAIF izpolnjuje pričakovanja** (strojno opremo zanemarimo). | ☑ |
| **Root Access** | Dostop do instance z administrativnimi pravicami. | ☑ |
| **Namestitev programja** | V instanco lahko namestimo potrebno programje. | ☑ |
| **Storage API** | Ločeni podatki vs. "user/scratch folder". | ☑ |
| **Workflow** | SLAIF pomeni resno izboljšavo glede na uporabo klasičnega HPC.<br>Bolj primeren pristop za običajnega razvijalca – NI Torque/Slurm. | ☑ |

# Hvala za pozornost!














Financerja / Financed by:

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA VISOKO ŠOLSTVO,
ZNANOST IN INOVACIJE

EuroHPC
Joint Undertaking

Projekt SLAIF: Slovenska tovarna umetne inteligence je finančno podprlo Ministrstvo za visoko šolstvo, znanost in inovacije. Projekt je bil na razpisu skupnega podjetja EuroHPC izbran za financiranje v okviru programov Obzorje Evropa ter Digitalna Evropa.

SLAIF: Slovenian AI Factory has been funded by the Ministry of Higher Education, Science and Innovation of Republic of Slovenia. At a call by EuroHPC JU, the project has received a positive funding decision under Horizon Europe and Digital Europe Programmes.


Get IT done.