TSP DATA

6.10.2026

Red Hat OpenShift AI je nasazený - co teď?



Miroslav Kolář
miroslav.kolar@tspdata.cz

Luděk Šafář
ludek.safar@prg.ai

scan qr below & win 👇

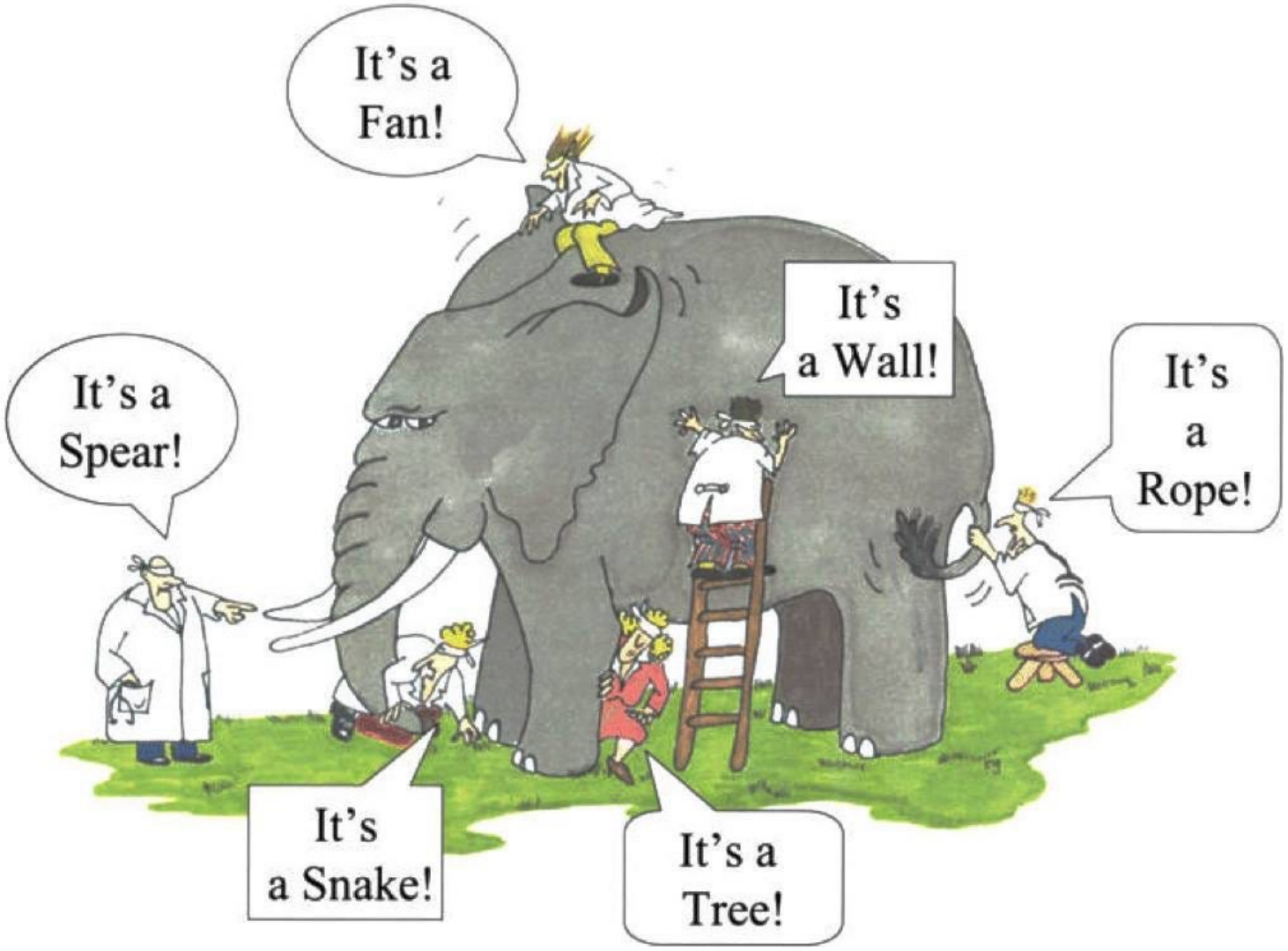


Namespace

Scénář

Zaplaceno, nasazeno.
Ale nic tam neběží.

-- ovšem nikdo nic nepokazil



Tři týmy

Infrastruktura	Aplikace	Data
Umí **postavit** platformu a **provozovat** ji.	Ví, jak **naplnit** business **požadavky**.	Dokáže **měnit** data na **hodnotu**.
U pokročilých komponent **potřebuje zadání** jak dál.	**Nemá** cluster-admin ani **chuť** platformu administrovat.	**Chce** si **stáhnout** knihovny, nástroje a modely zvenčí.
Bezpečnostní guidelines.	Hledá přesné zadání pro infra.	Hledá cestu, jak dostat veřejný model dovnitř.
Svoboda aplikačnímu týmu.	**Svobodu nedokáže využít.**	**Svobodu využije celou.**

Proč to v Kubernetes nebolelo

Dřív stačilo "jedno slovo".

Kubernetes:

"Dej mi namespace"

- 4 CPU, 8 GB RAM, 400GB PV, Postgres
- Aplikační tým to umí specifikovat
- Hranice = namespace

AI Platforma:

"Chceme skórovat klienty LLM"

- Které komponenty, jak dělit GPU, jaký runtime, guard rails
- Kompletní specifikaci nedokáže vyslovit nikdo
- Hranice = ?

Scénář všedního dne

- Volnost datovému týmu!
- Strategické rozhodnutí hostovat AI workloady interně.
- Nákup GPU (bez vazby na model a usecase).
- „Kdo to vlastní? „Asi infra…”

AI Platform Engineering

ODPOJENÉ PROSTŘEDÍ
oc-mirror a ImageDigestMirrorSet · zrcadlo balíčků · interní CA · import vah modelů

PLATFORMA
DSCInitialization · DataScienceCluster · které komponenty zapnout · šablony ServingRuntime · verze a upgrade okno

PUBLIKOVÁNÍ A KA
GatewayClass a Gate
ResourceFlavor · kvó

DETEKCE HARDWARU
Node Feature Discovery — operátor a jeho CR · Labely a taints na GPU nodech

icy · driver a KMM · MIG profily ·

Cohort ·
ResourceQuota, LimitRange

KOMPETENCE A PROCESY
Kdo zakládá projekt a nastavuje kvótu · kdo určuje verzi a kdy se upgraduje · kdo má mandát říct ne · kdo odpovídá compliance na retenci a audit · kdo onboarduje druhý a třetí tým · **kdo schvaluje, že model zvenčí smí dovnitř**

Jak se GPU reálně dělí

Dva pooly. Otázka je, kolik worklodů dáte do kterého.

	Dev a experimenty	Produkční inference
Mechanismus	Time-slicing	Dedikované GPU, nebo MIG
Izolace paměti	Ne	Ano
Proč to stačí / nestačí	Vývojář si notebook hlídá sám, spadlá úloha není katastrofa.	Spadlá inference je incident.

Cesta modelu dovnitř do firmy

externí registry → schválení → sken → podpis→ interní registry → nasazení

U images cesta existuje roky, má pravidla a každý ví, koho se zeptat.

externí registry → ? → ? → ?→ ? → nasazení

U vah modelu cesta neexistuje – model není image ani aplikace – a nikdo o tu cestu dosud nepožádal.

Pět otázek na zítra

1. Kdyby dnes aplikační tým chtěl nasadit model, kolik lidí musí něco udělat — a umíte je jmenovat?
2. Kolik komponent AI platformy máte zapnutých a kdo je vlastníkem každé z nich?
3. Jak se do vašeho prostředí dostane model zvenčí a kdo to schvaluje?
4. Kdo má mandát říct aplikačnímu týmu ne — a čím to zdůvodní?
5. Kdo u vás vlastní hardware profily a šablony ServingRuntime?

TSP DATA

Success driven by data.

Děkujeme za pozornost.

Miroslav Kolář
miroslav.kolar@tspdata.cz

Luděk Šafář
ludek.safar@prg.ai

scan qr below & win

JDEME LOSOVAT!

Losování právě teď na našem stánku

