

# Od datového skladu k datovým agentům

Jak AI mění přístup k datům

# Jak vlastně získáváme odpovědi z dat?

Co vše se skrývá na pozadí otázky:

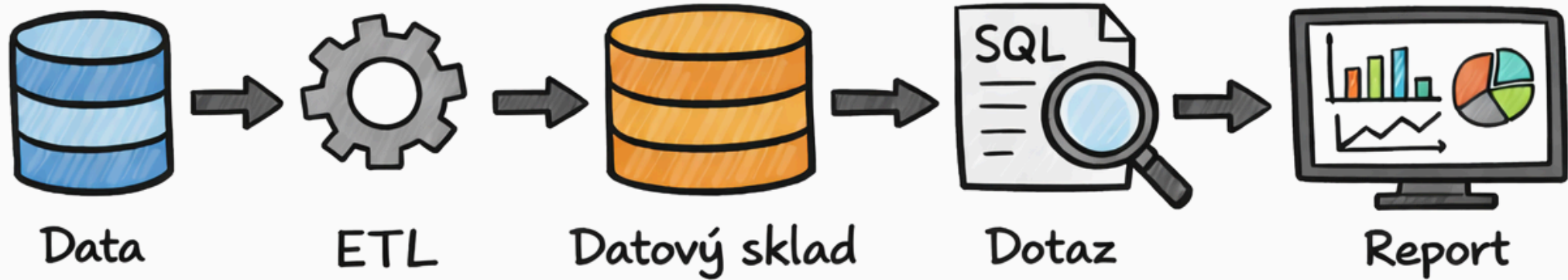
- „Kdo jsou naši nejaktivnější uživatelé za poslední týden?“

# Co se musí stát v systému?

- výběr dat
- spojení tabulek (JOIN)
- filtrování (čas, podmínky)
- agregace
- řazení

Jednoduchá otázka → složitý proces

# Klasický BI / DW pipeline



## Role v pipeline

- Data engineer → připravuje data
- Analytik → píše dotazy, sestavuje reporty
- Manažer → interpretuje výstup

## Vlastnosti klasického přístupu

- + strukturovaný
- + spolehlivý
- + opakovatelný
- ale náročný na znalosti

## Kde je problém?

### Přístup k datům

- nutnost znát schéma
- nutnost umět dotazovací jazyk (SQL)
- znalost business logiky

# Sémantická mezera

Člověk: uvažuje v konceptech jako `aktivní uživatel`

Systém: potřebuje přesně definované podmínky, formální dotaz

Co tedy znamená `Aktivní uživatel za posledních 7 dní` ?

- počet akcí?
- přihlášení?
- odehrané hry?

# Data ≠ informace

Data jsou uložena, ale význam není explicitně zmíněn poblíž dat.

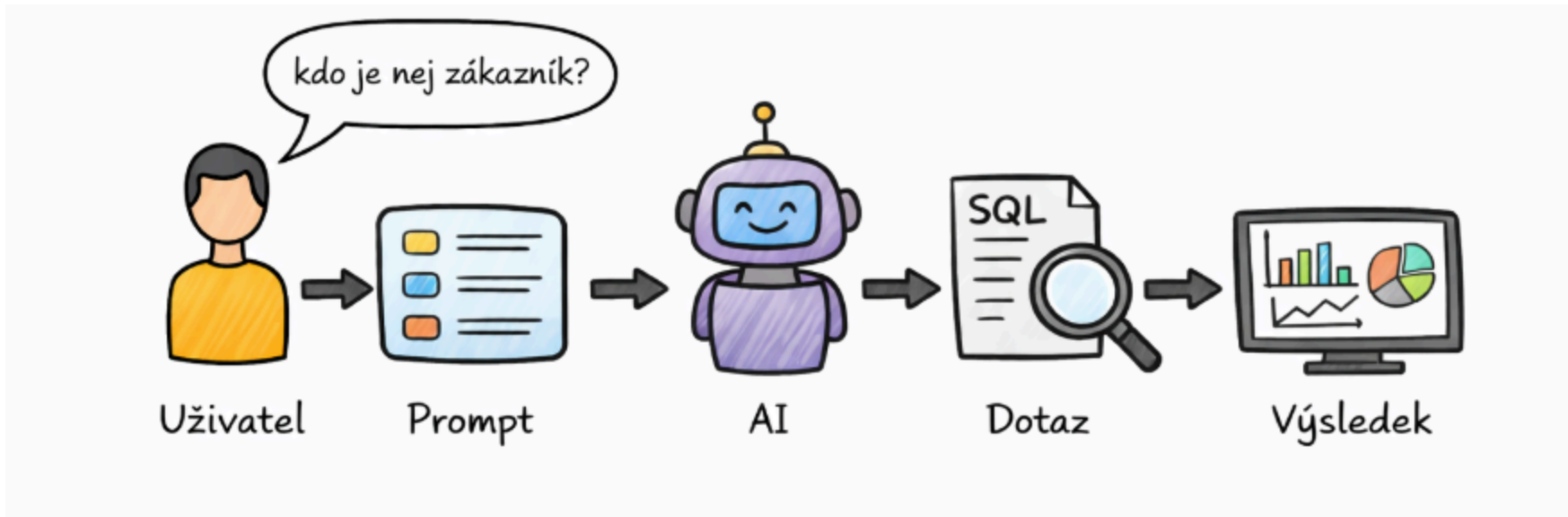
Je potřeba aby někdo převedl **fuzzy business význam** na **explicitní formální definici**

# Nový přístup: AI jako rozhraní

Co kdybychom se ptali přirozeným jazykem?



# Nový přístup



## Role AI:

- interpretuje záměr
- generuje SQL/Pandas/filtry/agregace

# Úrovně použití AI

- **Jednoduchý dotaz** (LLM, chatbot) k sestavení dotazu
- **Dotaz + dodání kontextu** uživatelem (schema)
- **Agentní přístup** (může si dohledat kontext, doptat se)

Každá úroveň přidává více schopností, ale i více komplexity a rizik

## Naivní LLM

Kdo jsou naši top zákazníci?

Nemá kontext, odhaduje strukturu, generuje **nějaký** generický SQL.

Vypadá správně... ale nemusí být!

Can you generate me an SQL query that will show:

Kdo jsou naši top zákazníci?

Jasně 😊 pojďme na to — otázka „Kdo jsou naši top zákazníci?“ se dá v SQL interpretovat několika způsoby, podle toho, co znamená *top*:

- 🏆 nejvyšší celková útrata
- 🛒 nejvíc objednávek
- 📦 největší objem nakoupených položek
- ⌚ nejaktivnější za určité období

Níže máš nejčastější variantu: **top zákazníci podle celkové útraty**.

### 💰 1. Top zákazníci podle obratu (nejčastější definice)

SQL

```
SELECT
  c.customer_id,
  c.customer_name,
  SUM(o.total_amount) AS total_spent
FROM customers c
JOIN orders o ON c.customer_id = o.customer_id
GROUP BY c.customer_id, c.customer_name
ORDER BY total_spent DESC
LIMIT 10;
```



# Schema-aware prompting

Prompt obsahuje schema -  
tabulky, sloupce, vztahy

## + Výhody

méně halucinací, lepší joiny,  
realističtější dotazy

## — Limity

stále nezná business logiku,  
jednorázový odhad

AI už „vidí strukturu“, ale stále  
nechápe význam

Can you generate me an SQL query based on this?

Tables:

customers(id, name)

orders(id, customer\_id, amount, created\_at)

Question:

Who are the top customers?

That's a classic question—but “top customers” can mean a few different things depending on your goal 😊 Let me give you the most common interpretations so you can pick what fits best.

### 🏆 1. Top customers by total spending (most common)

SQL

```
SELECT
  c.id,
  c.name,
  SUM(o.amount) AS total_spent
FROM customers c
JOIN orders o ON c.id = o.customer_id
GROUP BY c.id, c.name
ORDER BY total_spent DESC
LIMIT 10;
```



## Agentní přístup

Agentní přístup není jen o jednom dotazu a promptu, ale AI systém může **na základě svého vnitřního usouzení** volat další pomocné funkce, například:

- zjištění schematu tabulky
- ověření funkčnosti dotazu + oprava
- iterace na řešení
- upřesňující dotaz na uživatele

## Možný proces agentního systému

- 👤 se zeptá `kdo jsou naši top zákazníci`
- 🤖 zavolá funkci na **načtení schematu** databáze
- 🤖 **usoudí**, které entity a atributy jsou relevantní
- 🤖 **zeptá se uživatele** na upřesnění definice `top`
- 👤 popíše metriku
- 🤖 **vytvoří query**
- 🤖 **spustí query** a v případě chyby iteruje řešení problému
- 🤖 **vrátí výsledek**
- 👤 ověří, že výsledek odpovídá požadavku

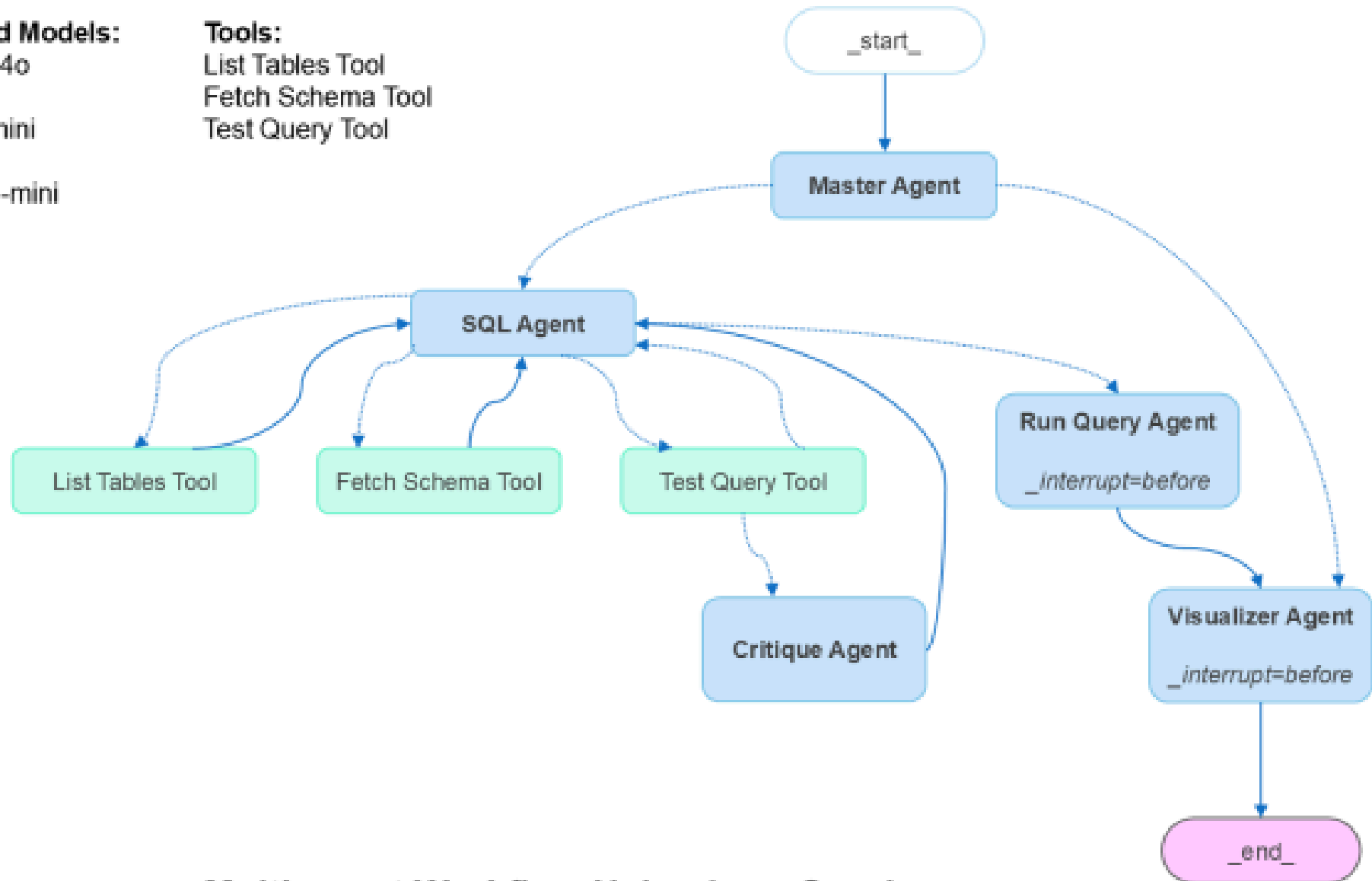
**Klíčová změna:** dotaz není jednorázový, ale je to plnohodnotný iterativní proces hledání odpovědi

### Agents and Utilized Models:

Master Agent- GPT-4o  
SQL Agent- o3-mini  
Critique Agent- o3-mini  
Run Query- o3-mini  
Visualizer Agent- o3-mini

### Tools:

List Tables Tool  
Fetch Schema Tool  
Test Query Tool



## Multi-agent Workflow Using LangGraph

# Posun paradigmatu díky Agentům

Klasický přístup:

- pevně definované dotazy pro dashboardy
- validované
- vysoká důvěra

AI přístup:

- dynamické dotazy
- flexibilita
- nutnost ověření a chápání kontextu uživatelem

Ideální hybrid využívající silné stránky obou - AI přístup pro exploraci, pak ukotvit a validovat klasicky



## Limitace Agentní přístupu

AI neřeší nejednoznačnost - odhaluje ji.

Co agent stále neví:

- co je „správná metrika“
- jaká data jsou validní
- jaké jsou business pravidla

# Co se děje uvnitř agenta?

Agent může dělat různé věci,  
komplikací je, že uživatel většinou vidí jen výsledek.

## Skrytá logika

- rozhodovací proces není viditelný
- mezikroky nejsou běžně dostupné
- debug logy bývají skryté

Běžný uživatel nástroje neví, *proč* agent došel k výsledku.

## SQL vs AI přístup

| SQL               | AI                 |
|-------------------|--------------------|
| vidím dotaz       | vidím výsledek     |
| kontroluji logiku | odhaduji správnost |
| přesný postup     | skrytý proces      |

Změna kontroly nad systémem a posun k **black box rozhodování**.

# Mapování úrovní použití AI na reálné nástroje

## Jednoduchý LLM

- ChatGPT, Copilot, ...
- generování SQL z textu

## Kontext + schema

- BI nástroje (Power BI, Tableau + AI)
- nástroje s přístupem ke schématu

## Agentní přístup

- systémy s přístupem k DB + iterací
- experimentální agenti (LangChain a podobné frameworky)
- interní firemní adhoc řešení

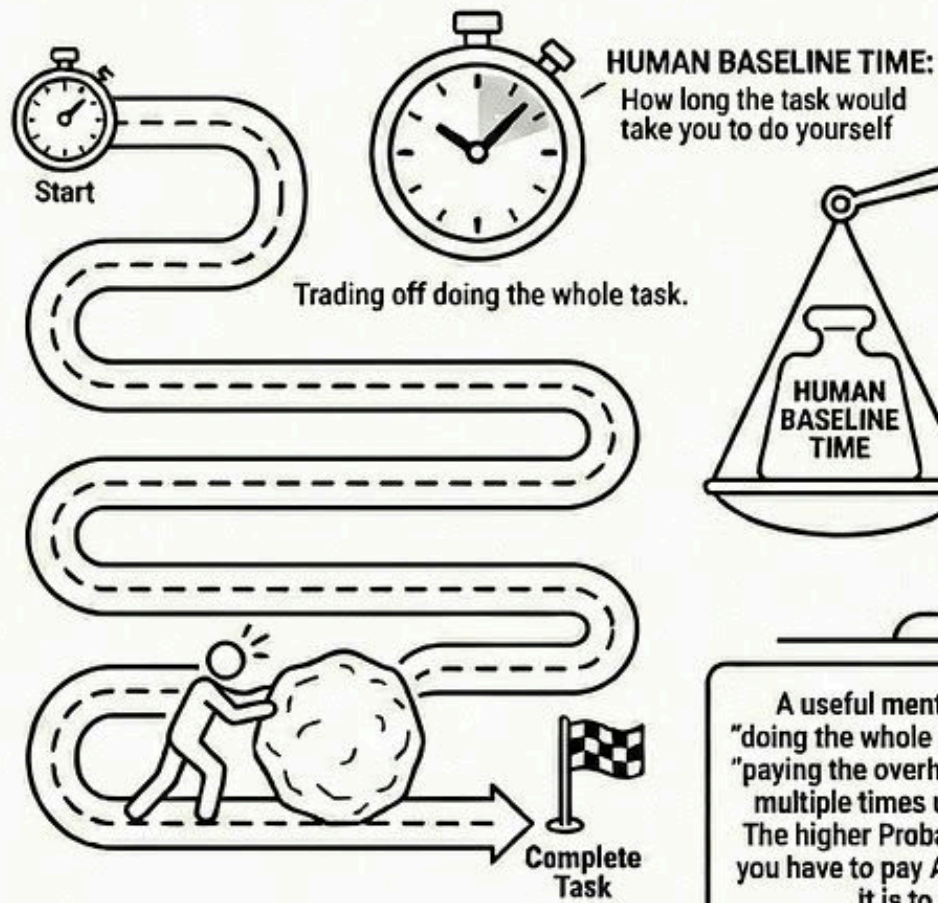
# Jak vyspělá je tato oblast?

- rychlý vývoj (2023–2026)
- nástroje jsou zatím nedokonalé
- velká variabilita kvality

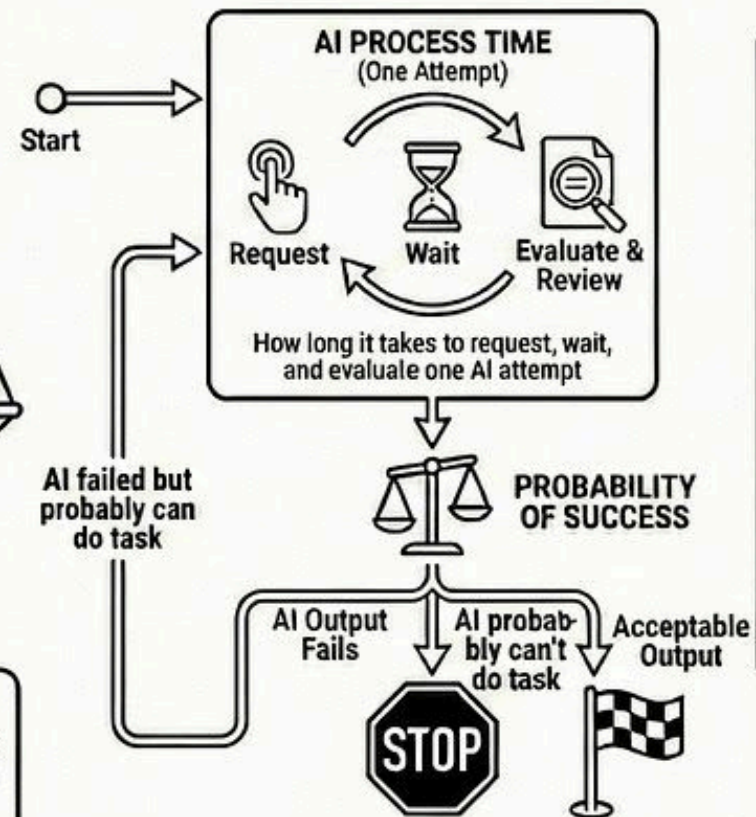
Neexistuje „standardní řešení“ kvůli rychlému zastarávání a velké konkurenci.

# DECIDING TO DELEGATE TO AI: THE TRADE-OFF

## OPTION 1: DOING THE WHOLE TASK (Human Baseline Time)



## OPTION 2: DELEGATING TO AI (Paying the Overhead Cost)



A useful mental model is that you're trading off "doing the whole task" (Human Baseline Time) against "paying the overhead cost" (AI Process Time) possibly multiple times until you get something acceptable. The higher Probability of Success is, the fewer times you have to pay AI Process Time, and the more useful it is to turn things over to the AI.

# Jakou odpověď od AI dostaneme?

## Běžné nástroje pro využití AI k dotazování

- jeden SQL dotaz
- jeden výsledek
- bez vysvětlení

Rychlé, ale zjednodušené. Efektivní pro integraci do dalšího workflow, protože odpověď je jednoznačná.

## Co by mohl dělat LLM

- nabídnout více interpretací
- vysvětlit rozdíly
- ukázat alternativy

Top zákazník může znamenat více věcí. Ale uživatel musí odpověď zpracovat a vybrat si.

## Zploštění odpovědi

- Nástroj vybere jednu interpretaci,
- ignoruje ostatní možnosti

Nuance se ztrácí. Uživatel si nemusí uvědomit "nejistotu" odpovědi a alternativy.



# AI a citlivá data

Použití SaaS řešení AI znamená:

- data (nebo jejich část) opouští systém
- dotazy mohou obsahovat citlivé informace
- poskytovatel může garantovat respektování soukromí a nepoužívání interních dat pro své účely, ale těžko prokazatelné

## Rizika

- únik dat
- porušení regulací (např. GDPR)
- nemožnost auditu

Problém bude zejména v oblastech jako: zdravotnictví, finanční sektor, interní firemní data.

# Lokální modely

Alternativa:

- model běží interně
- data neopouští infrastrukturu
- větší kontrola

## Trade-off

| Cloud AI       | Lokální model     |
|----------------|-------------------|
| vyšší kvalita  | nižší kvalita     |
| snadné použití | složitější provoz |
| riziko dat     | větší bezpečnost  |

# Jak poznáme správný výsledek?

AI nástroj vrátí:

- ✓ validní SQL
- ✓ validní výstup

Ale je to správně?

## Proč je to těžké?

- neexistuje jednoznačný „ground truth“
- business definice nejsou explicitní
- data mohou být nekvalitní

## Důsledek

Validace je na uživateli

## Jak ověřit výsledek?

- dává číslo smysl?
- odpovídá rozsahu dat?
- není extrémně velké / malé?
- lze výsledek spočítat jinak?

Jednoduché kontroly často odhalí chybu.

Je výhodné mít **širokou general knowledge**, protože pak snadněji padne chyba do oka.

# Typy chyb AI dotazů

## 1. syntaktická

SQL se nespustí

## 2. strukturální

špatný join, duplicity

## 3. logická

špatná metrika / agregace

## 4. interpretační

špatné pochopení zadání

Nejhorší typ chyby logická / interpretační - dotaz funguje a výsledek může vypadat ok  
(= těžké odhalit)

## Co se mění se zaváděním AI nástrojů?

- méně potřeby znát SQL syntax
- méně potřeby znát schéma

## Důležité omezení

- AI nezná vaše data
- kdo garantuje správnost?

AI může dát špatnou odpověď, která vypadá správně



## Datové modelování je stále klíčové

- kvalitní schéma → lepší výsledky
- jasné názvy → lepší interpretace
- čistá data → méně chyb
- dokumentace ke znalostní bázi je důležitá pro kontext (agent si ji rád přečte)

# Nová role uživatele

Dříve:

- psaní dotazů

Dnes:

- interpretace výsledků
- odhalování chyb
- chápání kontextu
- ověřování metodologie výpočtu

# Závěrečná myšlenka

AI usnadňuje přístup k datům (nejen skrze SQL), ale ztěžuje důvěru ve výsledek.

**Proč se vlastně o datech učíme?**

AI vám pomůže napsat dotaz, ale **nepozná, že je špatně**. To musíte umět vy