Big Data & AI

Jenda Hamáček

Proč?

- Učivo podle RVP
- Těžko uchopitelné, těžké a drahé na vyzkoušení
- Vede na aktuální témata jako sběr dat, zpracování dat AI, LLM

Jaká data existují

Pro začátek malinká

Jaká data existují

Pro začátek malinká

- Bit (0 nebo 1)

Jaká data existují

Pro začátek malinká

- Bit (0 nebo 1)
  - informace ano-ne - kamkoliv

Jaká data existují

Pro začátek malinká

- Bit (0 nebo 1)
  - informace ano-ne - kamkoliv
- Bajt - 8 bitů

Jaká data existují

Pro začátek malinká

- Bit (0 nebo 1)
  - informace ano-ne - kamkoliv
- Bajt - 8 bitů
  - jedno písmenko nebo malé číslo - zapamatuju si

Jaká data existují

Pro začátek malinká

- Bit (0 nebo 1)
  - informace ano-ne - kamkoliv
- Bajt - 8 bitů
  - jedno písmenko nebo malé číslo - zapamatuju si
- Kilobajty

Jaká data existují

Pro začátek malinká

- Bit (0 nebo 1)
  - informace ano-ne - kamkoliv
- Bajt - 8 bitů
  - jedno písmenko nebo malé číslo - zapamatuju si
- Kilobajty
  - textový email - Vytisknu na papír

Jaká data existují

Jaká data existují

- Megabajty

Jaká data existují

- Megabajty
  - fotka, kniha, sken, text můžu zpracovat v excelu

Jaká data existují

- Megabajty
  - fotka, kniha, sken, text můžu zpracovat v excelu
- Gigabajty

Jaká data existují

- Megabajty
  - fotka, kniha, sken, text můžu zpracovat v excelu
- Gigabajty
  - film, fotoalbum, aplikace (s grafikou), můžu uložit do databáze a zpracovat

Jaká data existují

- Megabajty
  - fotka, kniha, sken, text můžu zpracovat v excelu
- Gigabajty
  - film, fotoalbum, aplikace (s grafikou), můžu uložit do databáze a zpracovat
  - naprogramovat program, který to zpracovává (Python Pandas)

Jaká data existují

Větší data

Jaká data existují

Větší data

- Terabajty

Jaká data existují

Větší data

- Terabajty
  - jednotky: velikost disku - většinou filmy, fotky, aplikace

Jaká data existují

Větší data

- Terabajty
  - jednotky: velikost disku - většinou filmy, fotky, aplikace
  - desítky, stovky - spousty videí? (youtube - exabajty)

Jaká data existují

Větší data

- Terabajty
  - jednotky: velikost disku - většinou filmy, fotky, aplikace
  - desítky, stovky - spousty videí? (youtube - exabajty)
- Petabajty - 1024 Terabajtů

Jaká data existují

Větší data

- Terabajty
  - jednotky: velikost disku - většinou filmy, fotky, aplikace
  - desítky, stovky - spousty videí? (youtube - exabajty)
- Petabajty - 1024 Terabajtů
  - Ještě víc videí

Jaká data existují

Větší data

- Terabajty
  - jednotky: velikost disku - většinou filmy, fotky, aplikace
  - desítky, stovky - spousty videí? (youtube - exabajty)
- Petabajty - 1024 Terabajtů
  - Ještě víc videí
  - Záznamy chování lidí na internetu

Jaká data existují

Jaká data existují

- Exabajty - 1024 Petabajtů

Jaká data existují

- Exabajty - 1024 Petabajtů
  - Velikost Youtube řádově (ještě víc videí)

Jaká data existují

- Exabajty - 1024 Petabajtů
  - Velikost Youtube řádově (ještě víc videí)
- Zetabajty - 1024 Exabajtů

Jaká data existují

- Exabajty - 1024 Petabajtů
  - Velikost Youtube řádově (ještě víc videí)
- Zetabajty - 1024 Exabajtů
  - Velikost Internetu (64 ZB k roku 2020)

Co už označit za Big Data?

- Zpracováváme větší data než se vejdou do tradiční databáze? (několik Serverů?)
- Potřebujeme za běhu zpracovávat více dat než zvládne pár serverů? (Google vyhledávání, větší weby)
- Další Wiki charakteristiky: Rozmanitost, Spolehlivost, Hodnota, Kompletnost, ...

Kde se ty data berou?

Kde se ty data berou?

- Obsah ze sociálních sítí (videa, fotky, profily, propojení)
- Chování uživatelů sociálních sítí (scrollování, kliky, doba sledování)
- Všechny najeté jízdy Uberu (10 PB cluster v 2020)
- Všechna vyhledávání Google
- Záznamy o poloze chytrých telefonů

Kde se ty data berou?

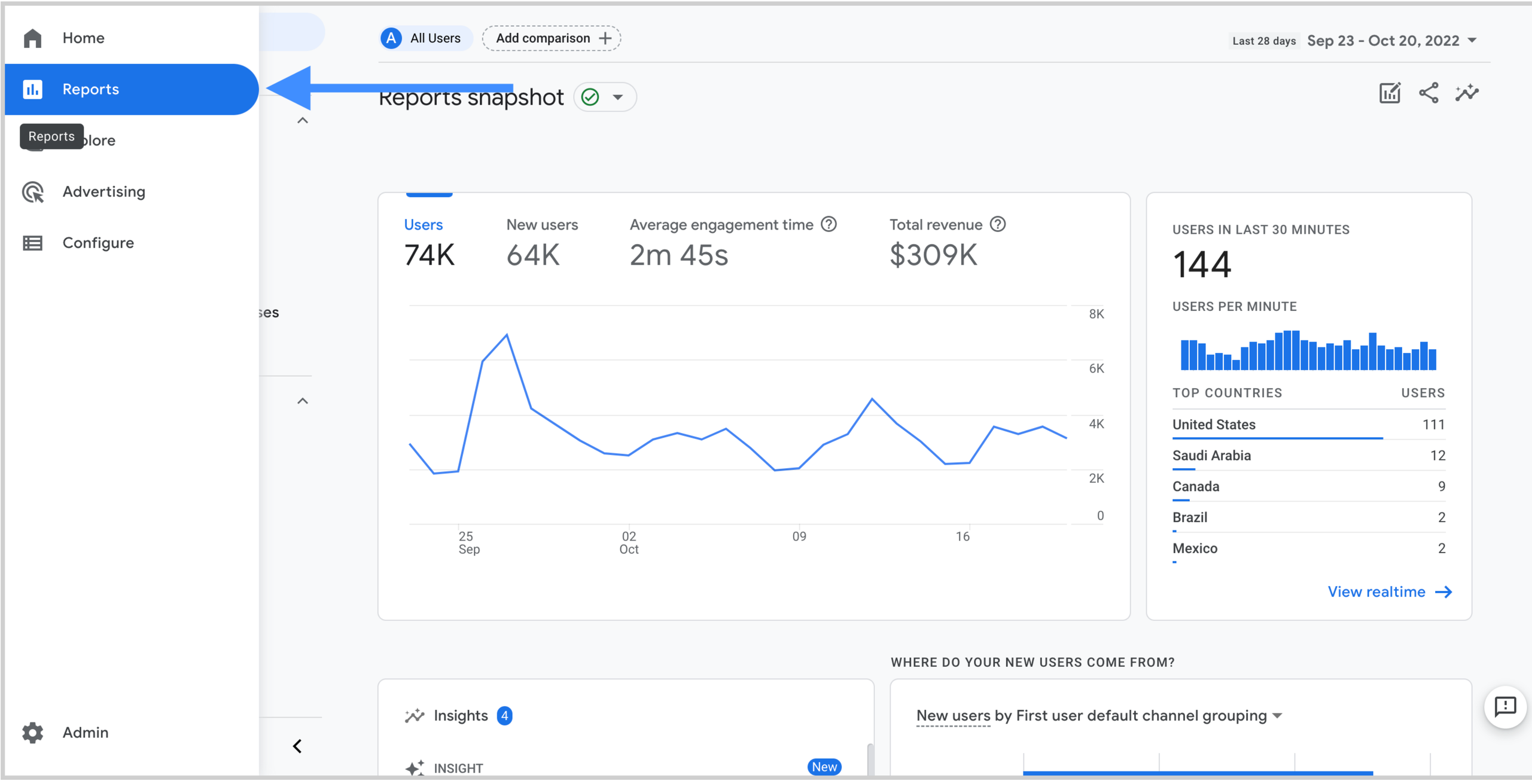
- Sebrané kliky na internetu (Avast 2020)
- Fotky odeslané robotickými vysavači pro analýzu
- Obecněji data sesbíraná ze všech IOT - světla, chůvičky, kávovary, televize, home assistenti...
- Data z novějších automobilů (diagnostika, kamery, poloha, používání) (Mozilla)

Příklad: podívejte se na sbíraná data na internetu

- 3rd party
  - Ukázka: gchd.cz , filtr google-analytics.com
- 1st party
  - Ukázka: google.com , filtr gen_204
  - Ukázka: search.seznam.cz , filtr v3
- Výjimky existují (např. duckduckgo)

Úkol

- Sbírá nějaká data o chování uživatelů web vaší školy?


Home
Reports
Reports
Explore
Advertising
Configure
Admin
All Users
Add comparison
Last 28 days
Sep 23 - Oct 20, 2022
Reports snapshot
Users
74K
New users
64K
Average engagement time
2m 45s
Total revenue
$309K
8K
6K
4K
2K
0
25
Sep
02
Oct
09
16
USERS IN LAST 30 MINUTES
144
USERS PER MINUTE
TOP COUNTRIES
USERS
United States 111
Saudi Arabia 12
Canada 9
Brazil 2
Mexico 2
View realtime
WHERE DO YOUR NEW USERS COME FROM?
Insights 4
INSIGHT
New
New users by First user default channel grouping

Jak to uložit?

- Zálohovací magnetické pásky
  - Vysoká kapacita (desítky TB), pomalý sekvenční přístup, levnější
- HDD
  - Vysoká kapacita (desítky TB), náhodný přístup, vyšší chybovost než pásky, dražší
- SSD
  - Nižší kapacita než HDD, dražší než HDD, rychlejší

Jak to zpracovat?

Schematický příklad Spark

- SELECT SUM(Length) FROM titles WHERE type = 'MOVIE'

1. Na každém serveru zvlášť Sum type = 'MOVIE'

2. Sum mezisoučtů

Jak to zpracovat?

- Deskriptivní analýza - nástroje podobné SQL
- Statistika
- Učení AI

AI

Příklady

AI

Příklady

- Je tento komentář urážlivý? (klasifikace, nlp, učení s učitelem)
- Jaké je riziko, že tento uživatel přestane používat naši službu? (regrese, učení s učitelem)
- Seskup naše uživatele podle podobnosti jejich chování (clustering)
- Logika řízení samořídícího auta (mnoho různých AI, reinforcement learning)
- Siri, Google Assistant - mj. syntéza řeči, převod řeči na text, pochopení dotazu
- Face ID - rozpoznání (obohaceného) obrazu
- Jaké je riziko, že tento člověk nesplatí půjčku? (regrese, učení s učitelem)

LLM

Generativní umělá inteligence

- Disclaimer: na výstup LLM se nikdy nedá 100% spolehnout!

LLM

Generativní umělá inteligence

- Disclaimer: na výstup LLM se nikdy nedá 100% spolehnout!
- To ale nemusí vždycky vadit

LLM

Generativní umělá inteligence

- Disclaimer: na výstup LLM se nikdy nedá 100% spolehnout!
- To ale nemusí vždycky vadit
  - Snadno ověřitelné úlohy (jednoduché programování)

LLM

Generativní umělá inteligence

- Disclaimer: na výstup LLM se nikdy nedá 100% spolehnout!
- To ale nemusí vždycky vadit
  - Snadno ověřitelné úlohy (jednoduché programování)
  - Manuální úlohy (převeď obrázek z tabule na textové zápisky)

LLM

Generativní umělá inteligence

- Disclaimer: na výstup LLM se nikdy nedá 100% spolehnout!
- To ale nemusí vždycky vadit
  - Snadno ověřitelné úlohy (jednoduché programování)
  - Manuální úlohy (převeď obrázek z tabule na textové zápisky)
  - Generování nápadů

LLM

Možnosti (zdarma)

- ChatGPT, OpenAI (US): https://chatgpt.com
- Claude AI, Anthropic (US): https://claude.ai/
- Mistral Large 2, Mistral (FR): https://chat.mistral.ai/chat
- Google gemini (US): https://gemini.google.com/app
- LLama 3.1 70b, Meta (US): https://www.meta.ai (není v ČR), ale https://huggingface.co/chat/ (dokonce bez přihlášení)
- Microsoft copilot: https://copilot.microsoft.com (jde i bez přihlášení)
- (Grok)

Síly LLM

- Obecná rada: model umí dobře to, co umí hodně lidí. Model umí špatně složité a okrajové věci, které umí jen málokdo.
- Práce s jazykem (přelož, přeformuluj, najdi gramatické chyby, ...)
- Programování (viz obecná rada)
- Zjednodušování, vysvětlování, sumarizace
- Pochopit otázku, kterou neumíme dobře položit
- Vyznat se v těžko čitelném obsahu (tahle chyba mi vypadla z PC, vysvětli co s ní, převeď mi citace těchto článků do takovéhohle formátu)
- Generování nápadů, postupů

Slabosti LLM

- Přímé počty (Kolik je 12322342 * 837529812?)
- Přiznání neznalosti.
- Velké (programovací) úlohy
- Komplexní logika, složité úlohy, úlohy o kterých existuje málo dat
- Rady s grafickým rozhraním
- Nejistota výsledku (včetně citací!)
- Přesná délka odpovědi

Příklady LLM

Kapitálka v html

Chtěl bych napsat html stránku s textovým obsahem. Text by měl začínat prvním písmenem, které je větší než ostatní a zbytek textu ho obtéká. Viděl jsem takový efekt v knize. Jak na to?

https://chatgpt.com/share/d22e0d90-3f7e-4452-940f-9fc024c4cb8a

Příklady LLM

Souhrn "dopis před odjezdem"

Tu je dopis "před odjezdem" pro účastníky akce, kam jedu. Co je potřeba si s sebou vzít nebo připravit?

- https://chatgpt.com/share/36147895-80b9-475c-81c4-83c1295bc4f6

Příklady LLM

Nachystejte potřebné znalosti LLM

- https://chatgpt.com/share/cbb6bc8c-4d1b-4da7-871d-d197b9988ee3
  - Pokud nevíte jak se zeptat, zeptejte se na to
  - Oddělujte kód od zadání (ideálně ```, případně cokoliv jako třeba ---)
  - Pamatujte, že i to co vypíše model dokáže při příštím dotazu využívat
  - Nebojte se ptát se na vysvětlení. Můžete upřesnit úroveň vyjadřování (vysvětluj pro dítě na základní škole, jsem učitel co pracuje s SQL poprvé)
  - Vkládejte celé chybové hlášky s popisem kde se vzaly.

Příklady LLM

Nápady

- I am doing sql in person course. I am giving my students tasks solvable by SQL SELECT. After a few tasks, I would like to ask them a question if they understand the concepts via some online tool. Can you give me some suggestions for such tools?

- https://chatgpt.com/share/2a99f85f-c722-4ff1-8fb9-226452799ec6

Tipy LLM

- Modely bývají chytřejší v angličtině. Pokud to zvládáte, doporučuji se ptát anglicky.
- Dnešní modely zvládají vstup o velikosti 128k tokenů. Tj. anglicky řádově necelá kniha, česky sto stránek textu.
  - Tj. nebojte se psát dlouhé zadání a vkládat dlouhé kusy textu, články apod.
- Přesto při dlouhém chatu občas zapomínají začátek.
  - Optimalizují tak cenu generování.

- To je vše, tádadádadá