Junior Data Scientist Akadémia 3. alkalom

- Adatgyűjtés
- Bash

Adatgyűjtés: "Garbage in, garbage out..."

->6 lépés -> 1, Adatgyűjtés -> Itt bukhat el.

2, Adattárolás

3, Adattisztítás

4, Adatelemzés

5, Vizualizálás -> Itt bukhat el.

6, Döntés -> Itt bukhat el.

Google Tag Manager -> követőkódot tud kezelni

Felelős -> adatminőségért

Mit gyűjtsünk? -> Vannak korlátok: -> nem az adattárolás drága

-> "nem jutott eszembe"

-> etikai, jogi korlátok

-> programozói források hiánya

-> oldal betöltési sebesség (weblapoknál)

- Hogyan döntsd el, hogy mit akarsz gyűjteni? -> menj végig a termékeden

-> gondolkozz visszafelé (1-on-1-ek)

=> Workshop! Workshop! Workshop!

Hogyan működik?

Website -> Tracking scripts => Data warehouse

↓
Frontend => Production server -> termék

=> Data saver -> Data server -> akár több szerver is lehet

Adatgyűjtési sémák -> SQL -> strukturált adatok
-> noSQL -> nem strukturált adatok
-> Json -> általános dinamikus lista
-> graph adatbázis rendszer -> Pl: neofog, neo4j

Bash:

-> nem adatelemző nyelv

-> ez egy magasító a nyelvek közt

-> direkt központi nyelv / hagyom központi / alapértő nyelv

-> Fájlrendszer

-> Data Science at the Command line (könyv)

-> SSH => bejelentkezés

clear -> kiseditja a képernyőt

cd.. -> felfelé lép

ls -> listáz

pwd -> Hol vagy?

cd -> visszalépés az elejére

Command prompt.

- grep lgrep lakasok.csv -> mondatokat keres

- head -100 lakasok.csv | wc -c
- cut -d',' -f3,4,11 lakasok.csv > oszlopok.csv
- cut -d',' -f4 lakasok.csv | cut -d'_' -f1 > varos.csv
- grep "Volkswagen" lakasok.csv | head
- cut -d',' -f11 lakasok.csv | sort -n
- sort -n -t',' -k11 lakasok.csv | tail

- sort -u -> duplikációkat tünteti el
- uniq -> összegyűjti sorokat

- cut -d',' -f2 lakasok.csv | sort | uniq -c
- cut -d',' -f2 lakasok.csv | sort -u | grep -v 'city' | grep -v 'UNKNOWN' | wc -l

- Script-írás és automatizálás
  - mcedit .sh
  - crontab -e

- psql -U user14 -d postgres -f proba.sql -F','

Kérdések:

- Adatgyűjtés
  1. Milyen adatokat gyűjtenél le?
  2. Milyen paramétereket?
  3. Milyen struktúrában?

Tényadatok

Eseményadatok

Vezetői szempontú változó adatok

Autó műszaki állapot

Star és Snowflake séma