FAKULTA RIADENIA A INFORMATIKY

ŽILINSKÁ UNIVERZITA V ŽILINE

ANALÝZA DÁT DOPRAVNÝCH NEHÔD

SEMESTRÁLNA PRÁCA Z PREDMETU DATABÁZOVÁ ANALYTIKA A VÝKONNOSŤ

Autor: **Bc. Matej Poljak**

Učiteľ: **doc. Ing. Michal Kvet, PhD.**

Akademický rok: **2024/2025**

Obsah

[Cieľ semestrálnej práce 3](#_Toc191677126)

[Popis dát 3](#_Toc191677127)

[Transformácia dát do databázy 3](#_Toc191677128)

[Databázová analytika 5](#_Toc191677129)

[Prehľad o počtoch dát 5](#_Toc191677130)

[Top mesiac/e pre každý kraj podľa počtu umrtí za rok 2024 6](#_Toc191677131)

[Optimalizácia výkonnosti 6](#_Toc191677132)

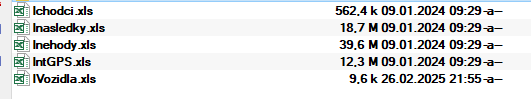
[Záver 6](#_Toc191677133)

# Cieľ semestrálnej práce

Cieľom práce bolo získať dáta, nad ktorými následne vykonáme analýzu. Touto analýzou zistíme zaujímavé informácie vyplývajúce z dát a tak preukážeme schopnosť analytického pohľadu na ne. Na to sme mali využiť analytické a agregačné funkcie, ktoré platforma Oracle poskytuje. Okrem toho sme sa mali pokúsiť optimalizovať rýchlosť dopytu dát s využitím indexov.

# Popis dát

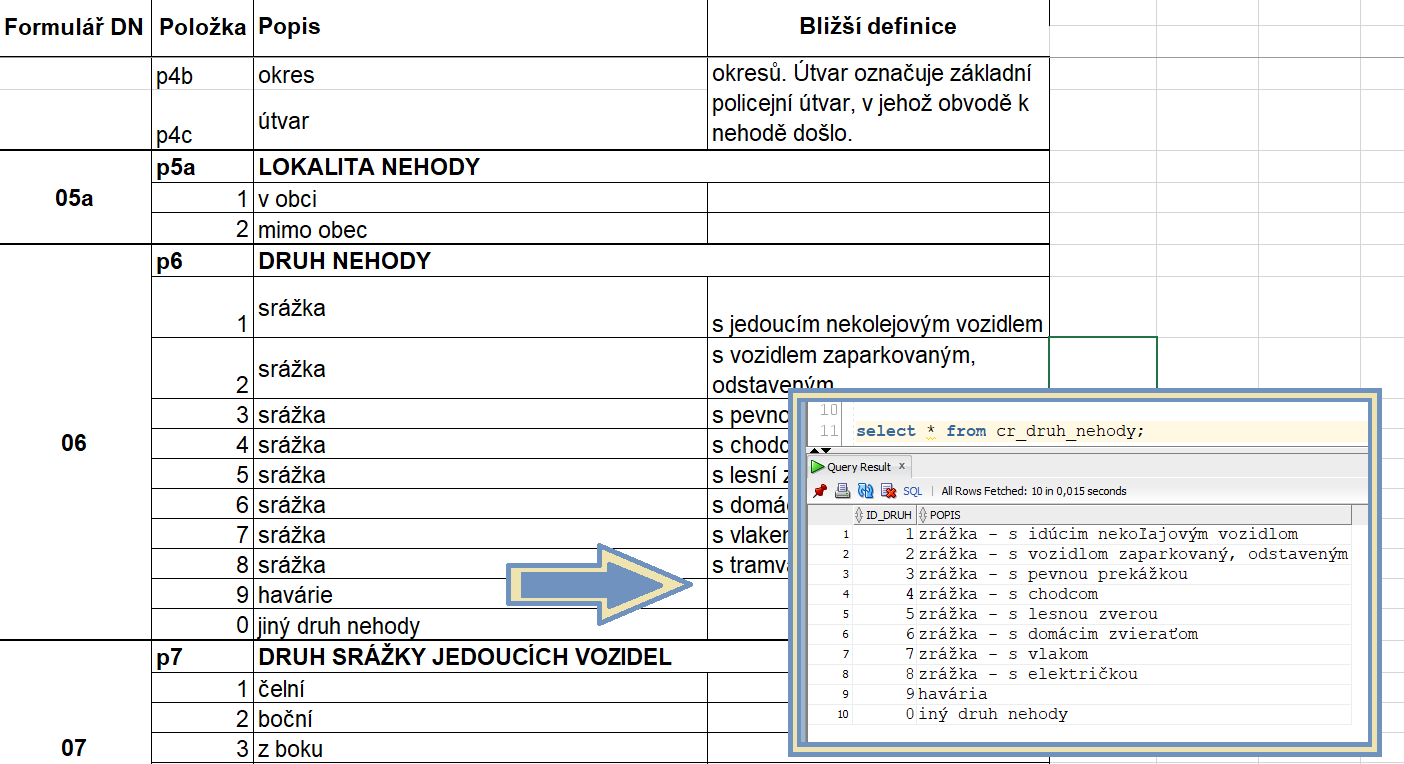
Na prácu som si zvolil údaje o dopravných nehodách v Českej republike, ktoré poskytuje česká polícia na mesačnej báze. Dáta môžeme získať zo stránky [*https://policie.gov.cz/clanek/statistika-nehodovosti-900835.aspx*](https://policie.gov.cz/clanek/statistika-nehodovosti-900835.aspx), kde si zvolíme rok, z ktorého dáta požadujeme. Stiahneme RAR zložku, ktorá obsahuje súbory s príponou *xls*. Súbory sú rozdelené na záznamy o nehodách, o vozidlách spojených s nehodou, o chodcoch spojených s nehodou a záznamy o následkoch dopravných nehôd (obrázok 1). Ja som si zvolil pre ďalšie spracovanie dáta o nehodách všeobecne a dáta o vozidlách, čiže dva typy *xls* súborov, pričom som zobral dáta z 3 rôznych rokov: 2023, 2024, 2025.



Obrázok 1 – obsah stiahnutej RAR zložky

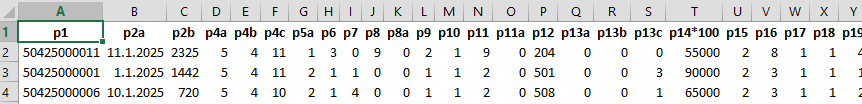
# Transformácia dát do databázy

Na stránke, uvedenej v predchádzajúcej kapitole, som si stiahol aj dokument *Položky\_formuláře\_WEB.xlsx* (príklad zobrazený na obrázku 2), ktorý obsahuje význam a hodnoty pre jednotlivé stĺpce, ktorými sú naplnené záznamy o nehodách a vozidlách zúčastnených pri nehode. Na základe týchto popisov som si pre každý referencujúci stĺpec vytvoril databázovú tabuľku, ktorá obsahuje významy jednotlivých ID hodnôt, na ktoré sa záznamy o nehodách a vozidlách odkazujú.



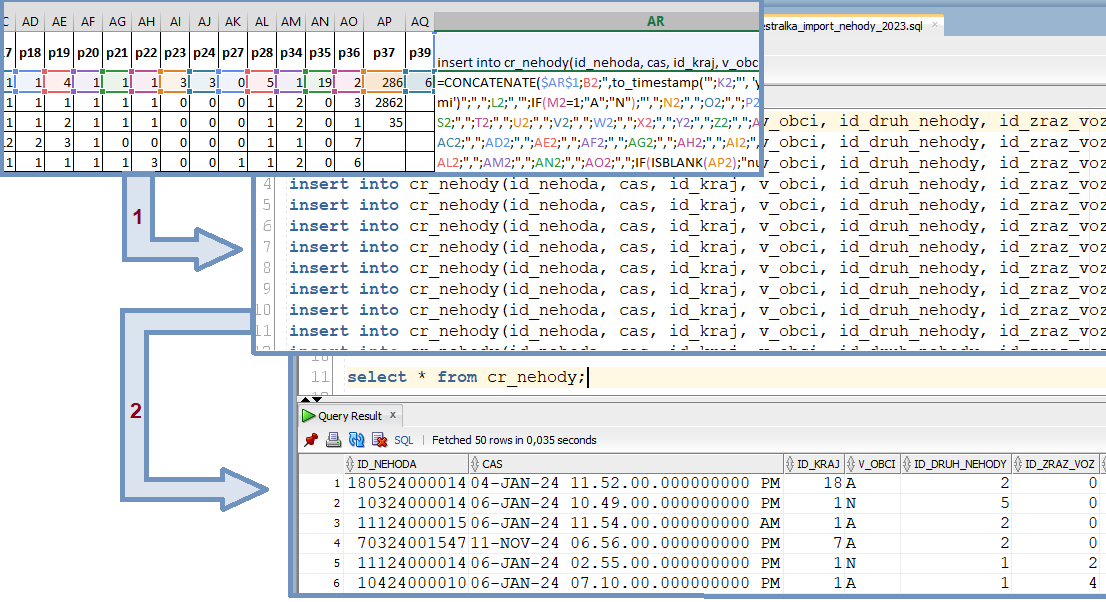
Obrázok 2 – popis hodnôt stĺpcov, na ktoré sa záznamy referencujú

Potom som si vytvoril databázovú tabuľky *cr\_nehody* a *cr\_vozidla* podľa štruktúry jedného riadku súboru *lnehody.xls*, resp. *lvozidla.xls*. Takýto záznam je ukázaný na obrázku 3.



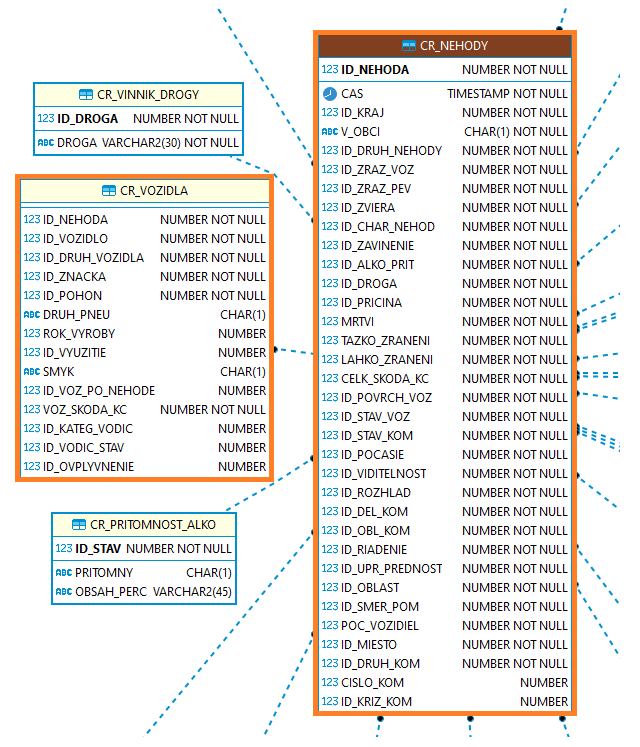
Obrázok 3 – ukážka záznamov o nehodách v xls súbore

Následne som mohol naplniť tabuľky o nehodách a vozidlách hodnotami zo súborov. Najprv som si napísal príkaz, ktorý mi v *xls* súbore vygeneroval sql príkaz *insert*. Všetky vytvorené sql príkazy som si uložil do súboru a potom som spustil vykonanie skriptu v nástroji *SqlDeveloper*. Tento postup je znázornený na obrázku 4.



Obrázok 4 – transformácia záznamov do databázy

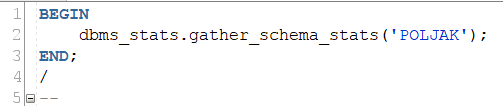
Pre prácu teda budeme využívať (okrem všetkých tabuliek s popismi) databázovú reprezentáciu záznamov zodpovedajúcu obrázku 5.



Obrázok 5 – dátový model najpodstatnejších tabuliek zvýraznených oranžovým rámikom

# Databázová analytika

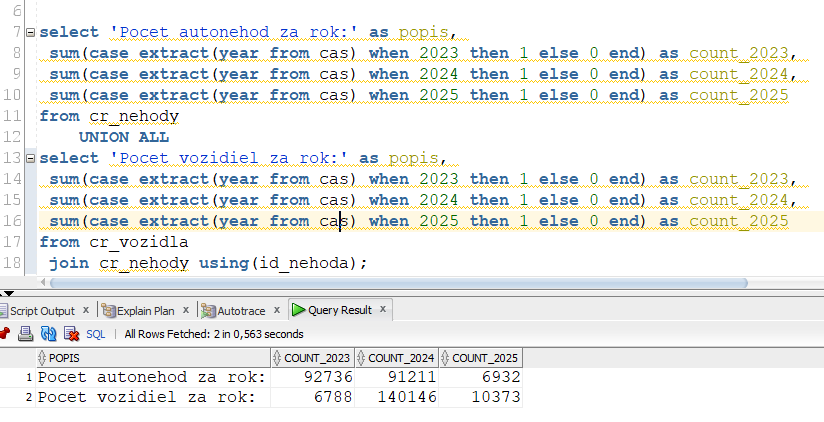
V tejto časti si ukážeme zopár zaujímavých informácií, ktoré sme získali databázovým príkazom *select* spolu s využitím analytických a agregačných funkcií. Pred analýzou výkonnosti si najprv nad všetkými tabuľkami, ktoré prislúchajú k schéme „*POLJAK*“ pregenerujeme štatistiky, pretože sme vložili veľké množstvo nových dát (obrázok 6).



Obrázok 6 – pregenerovanie štatistík pre aktualizáciu kvôli novým tabuľkám

### Prehľad o počtoch dát

Na úvod si pre lepšiu predstavu zistíme najzákladnejšiu štatistiku a to počty záznamov o nehodách a vozidlách za jednotlivé roky. Túto informáciu vieme pekne zobraziť pomocou pivotovej transformácie dosiahnutej použitím agregačnej funkcie *sum*, ktorá pripočíta do stĺpca záznam prave vtedy, keď spĺňa kritérium príslušnosti roku, pre ktorý je daný stĺpec určený. Stĺpec *cas* je *NOT NULL*, preto nemusíme ošetrovať podmienku s chýbajúcou hodnotou. Takéto dopyty sme vytvorili nad oboma tabuľkami nehôd aj vozidiel a následne sme ich zjednotili pre ucelenú informáciu v jednom výpise. Skript aj výpis je zobrazený na obrázku 7.



Obrázok 7 – počty nehôd a vozidel za jednotlivé evidované roky

### Top mesiac/e pre každý kraj podľa počtu umrtí za rok 2024

Prvou informáciou, ktorá nás môže zaujímať, je koľko a v akých mesiacoch pre konkrétny kraj Českej republiky (ďalej ČR) bol zaznamenaný najvyšší počet úmrtí spôsobených dopravnou nehodou za vymedzené obdobie rok 2024.

# Optimalizácia výkonnosti

# Záver