# Projekt (ENGETO Datová akademie): Data o mzdách a cenách potravin a jejich zpracování pomocí SQL

#### Richard Kosina

Zadáním projektu bylo **připravit robustní datové podklady**, ve kterých bude možné vidět **porovnání dostupnosti potravin na základě průměrných příjmů za určité časové období**.

Jako dodatečný materiál bylo nutné připravit také tabulku s HDP, GINI koeficientem   
a populací **dalších evropských států** ve stejném období, jako primární přehled pro ČR.

Primárními podklady pro přípravu datového podkladu týkajícího se vývoje cen, mezd a HDP byly tabulky „czechia\_payroll“, „czechia\_price“ a „economies“ z ENGETO studijní databáze (podklady pro tabulky jsou data z otevřených datových zdrojů). Prvním podstatným krokem je, si nejdříve všechny tři tabulky pořádně projít a zjistit strukturu dat.

Rozhodl jsem se, že si nejdříve vytvořím dvě tabulky týkající se cen a mezd. Na tabulku „czechia\_price“ jsem skrze JOIN připojil tabulku „czechia\_price\_category“, abych dostal   
do tabulky konkrétní názvy kategorií potravin. Skrze agregační funkce, jsem si vytvořil z datum měření, roky a čtvrtletí, aby mi to korespondovalo se strukturou plánovaně připojené tabulky vycházejících s tabulky „czechia\_payroll“. Podstatný poslední krok poté bylo využití GROUP BY funkce, kdy jsem si vizualizované sloupce nechal shluknout dle názvu potravin, roku a následně čtvrtletí měření. Tím jsem získal přehled o ceně základních potravin od 1. čtvrtletí roku 2006 do   
4. čtvrtletí 2018, postupně a přehledně pod sebou. Důležitým krokem také bylo agregování průměrné ceny potraviny (samozřejmě včetně zaokrouhlení), čímž dostávám ve výsledném náhledu zprůměrovanou cenu potraviny za všechny regiony a data měření v daném čtvrtletí daného roku.

Poté bylo nutné vytvořit si druhou tabulku vycházející z „czechia\_payroll“. Na tabulku „czechia\_price“ jsem skrze JOIN připojil tabulku „czechia\_payroll\_industry\_branch “, abych dostal do tabulky konkrétní kódy pro jednotlivá odvětví ekonomiky. Skrze rok, jsem připojil k tabulce „czechia\_payroll“ také tabulku economies“, abych dostal údaje o výši HDP v České republice   
v požadovaném roce. V zobrazení jsem si vydělením výše HDP a počtu obyvatel, vyjel také sloupec o výši HDP v daném roce na jednoho obyvatele. Údaje o HDP jsou k dispozici jen za rok, tudíž   
se tyto údaje připojili ke každému řádku u odvětví a čtvrtletí v daném roce. Kromě základních údajů, co jsem si nechal zobrazit (název, kód odvětví, průměrná mzda, rok, čtvrtletí, výše HDP), bylo nutné v dalším kroku si také určit upřesňující podmínky, jejichž parametry jsem zjistil   
z tabulek „czechia\_payroll\_unit „, „czechia\_payroll\_value\_type „ a „czechia\_payroll\_calculation „. Bylo tedy nutné, abych si nechal zobrazit jen řádky s informací o průměrné mzdě v odvětví (aby   
se mi nezobrazovalo počet zaměstnanců), přepočteném stavu průměrné mzdy (abych se dostal   
na údaje včetně příplatků, přesčasů, nemocenské a dovolené) a taky mezi lety 2006 a 2018 (aby mi to korespondovalo s daty, které jsou k dispozici pro zobrazení v tabulce „czechia\_price“). Podstatný poslední krok poté bylo využití GROUP BY funkce, kdy jsem si vizualizované sloupce nechal shluknout dle kódu odvětví, roku a následně čtvrtletí měření.

Práce s daty byla extrémně náročná na počítání, tudíž napojení takto dvou vytvořených tabulek jsem musel udělat skrze VIEW funkci, kdy jsem si obě tabulky nechal při použití funkce VIEW zobrazit a až tyto dvě VIEW tabulky „ceny“ a „platy“ jsem poté připojil k sobě. Ušetřilo mi to neskutečný množství času, když jsem to zkoušel spojovat k sobě klasicky, nemohl jsem se časově dopočítat.

Využil jsem klasický JOIN, aby mi každý údaj v tabulce „platy“ se připojil k patřičnému řádku   
v tabulce „ceny“, tudíž aby každý údaj v daném roce a čtvrtletí u dané potraviny se do řádku spároval s každým řádkem o údaji ve čtvrtletí o průměrné mzdě a odvětví. Nechtěl jsem žádné nulové hodnoty, který by vzniky využitím LEFT JOIN. Připojoval jsem obě tabulky skrze roky   
a také čtvrtletí. JOIN obou VIEW jsem spojil s vytvořením tabulky, tudíž mi pak vznikl samostatný datový podklad, využitelný pro další analýzu a zodpovězení výzkumných otázek. Tabulku jsem dle zadání pojmenoval „t\_richard\_kosina\_project\_SQL\_primary\_final“.

Vytvoření sekundárního datového podkladu s výši HDP, GINI a populací dalších evropských států (tedy mimo Českou republiku) vychází z tabulek „countries“ a „economies“. Zobrazení došlo   
z tabulky „economies“ a to názvu země, roku, výše HDP, výše GINI koeficientu a výše populace. Vytvořil jsem si také přepočet HDP na obyvatele. Tento údaj nesedí u Estonska, kde je zřejmě špatně uvedený údaj o celkovém HDP za rok (Estonsko by mělo mít hodnoty minimálně o řád vyšší), údaje u ostatních zemí zhruba odpovídají v porovnání s dalšími dostupnými zdroji. Důležité bylo připojit tabulku „countries“ (jednoduše skrze název země) za využití LEFT JOIN (počítal jsem totiž s nulovými hodnotami u parametrů z „economies“, které se netýkají konkrétních zemí). Posledním krokem poté bylo skrze klauzuli „WHERE“ upřesnit podmínky zobrazení, tady jsem využil připojenou tabulku „countries“ pro podmínku zobrazení jen zemí z kontinentu Evropa, kde není hlavní město Praha (abych neměl v přehledu Českou republiku, tu mám již v primárním přehledu) a léta zobrazení od roku 2006 až do roku 2018 (abych měl srovnatelné období s primární tabulkou). Tento proces jsem použil včetně vytvoření požadované tabulky, kterou jsem dle zadání projektu pojmenoval „t\_Richard\_Kosina\_project\_SQL\_secondary\_final“.

Odpovědi na výzkumné otázky jsem analyzoval jen z primárního podkladu, který mi stačil   
na zodpovězení všech výzkumných otázek. Níže se budu věnovat již jen interpretaci výsledných dat. K vytvoření datových podkladů pro odpovědi, jsem využil především WINDOW fuknci LAG, která mi v jednom řádku umožnila zobrazení jak údajů za daný rok, tak také za předchozí rok. Mohl jsem pak tedy vypočítat meziroční změnu za využití GROUP funkce (dle názvu požadované kategorie a roku). Jedinou nevýhodu bylo to, že pokud jsem si zobrazil více kategorií pod sebou (nevyužil jsem WHERE klauzuli, která mi zobrazí jen konkrétní potravinu/odvětví), měl jsem   
v roce 2006 (první rok měření, ve kterém ale nelze analyzovat změnu parametru), ve sloupci o údaji z předchozího roku, údaj z předchozího řádku (jak je to správně), který se týkal předchozí kategorie. Pro analýzu to ovšem nemělo vliv, protože v prvním roce měření nebylo možné, změny analyzovat.

1. **Rostou v průběhu let mzdy ve všech odvětvích, nebo v některých klesají?**Mezi lety 2006 a 2018 stoupla mzda ve všech odvětvích a to o desítky procent. Nejvýrazněji mzdy stouply v odvětví zdravotní a sociální péče o 78 % a zpracovatelském průmyslu o 72 %. Nejnižší nárůst mezi lety 2006 a 2018 byl naopak u odvětví peněžnictví a bankovnictví o 37 %. Meziroční růst mezd byl průběžně zaznamenaný u všech odvětvích, výjimkou byly rok 2009-2013 v době, kdy v České republice byla hospodářská recese a v těchto letech u některých roků, probíhalo u některých odvětví meziroční snížení průměrného výdělku. Nejvýraznější meziroční pokles lze sledovat v roce 2013 o 9 % u odvětví peněžnictví a bankovnictví. Nejvýraznější meziroční nárůst lze ve sledovaném odvětví pozorovat v roce 2008 u odvětví těžby a dobývání o 14 %. Odvětví těžby a dobývání je také jediné odvětví ve kterém meziročně klesla mzda mimo roky 2009-2013. Stalo se tak v letech 2014 a 2016, kdy došlo k v obou letech k meziročnímu poklesu o 1 %.
2. **Kolik je možné si koupit litrů mléka a kilogramů chleba za první a poslední srovnatelné období v dostupných datech cen a mezd?**V prvním čtvrtletí roku 2006 bylo možné si za průměrný měsíční výdělek koupit 1 334 kg chleba, zatímco ve čtvrtém čtvrtletí 2018 to bylo již 1 463 kg chleba (nárůst o 9,6 %). V případě mléka je tento nárůst ještě vyšší, v prvním čtvrtletí roku 2006 bylo možné si za průměrný měsíční výdělek koupit 1 430 l mléka, zatímco ve čtvrtém čtvrtletí roku 2018 to již bylo 1 848 l mléka (nárůst o 29,2 %).
3. **Která kategorie potravin zdražuje nejpomaleji (je u ní nejnižší percentuální meziroční nárůst)?**Mezi lety 2006 a 2018 došlo k nejnižšímu meziročnímu nárůstu u krystalového cukru, kdy mezi těmito lety jako u jediné kategorie potravin došlo dokonce k poklesu ceny o 27 %. Cukr ovšem v průběhu let mění cenu dynamičtěji (například v roce 2011 došlo k meziročnímu nárůstu ceny o 22 %). Nejmenší meziroční nárůst cen lze tak vypozorovat u vepřové pečeně s kostí a polotučného mléka, kde nejvyšší meziroční nárůst ceny lze pozorovat v roce 2017 o 8 % u vepřové pečeně, a 2011 a 2014 u polotučného mléka.
4. **Existuje rok, ve kterém byl meziroční nárůst cen potravin výrazně vyšší než růst mezd (větší než 10 %)?**Ne, takový rok ve sledovaném období neexistuje. Jediný rok, kdy byl růst cen potravin oproti růstu mezd výraznější je rok 2013, kdy cena potravin oproti růstu mezd byla meziročně vyšší o 7 %.
5. **Má výška HDP vliv na změny ve mzdách a cenách potravin? Neboli, pokud HDP vzroste výrazněji v jednom roce, projeví se to na cenách potravin či mzdách ve stejném nebo následujícím roce výraznějším růstem?**Pokles HDP byl pozorován v letech 2009 (pokles o 5 %) a a 2012 (pokles o 1 %). V následujícím roce 2013 došlo jako v jediném z pozorovaných let k poklesu výše průměrné mzdy (pokles o 2 %). Jinak meziroční vliv růstu HDP na výši růstu mezd a cen není příliš průkazný, většinou roste HDP pomalejším způsobem, jak rostly ceny a mzdy. Jediný rok kdy lze sledovat vliv růstu HDP (růst o 5 %) je rok 2017, kdy v roce 2017 průměrná mzda stoupla o 6 % a v následujícím roce o 8 %.