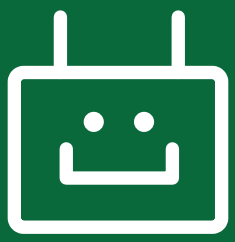
Microsoft a czechitas #NováGenerace uvádí

DATA

...úvod do datové analytiky

Tutoriál pro první kroky s daty

Tento tutoriál je vhodný pro začátečníky od 12 let.



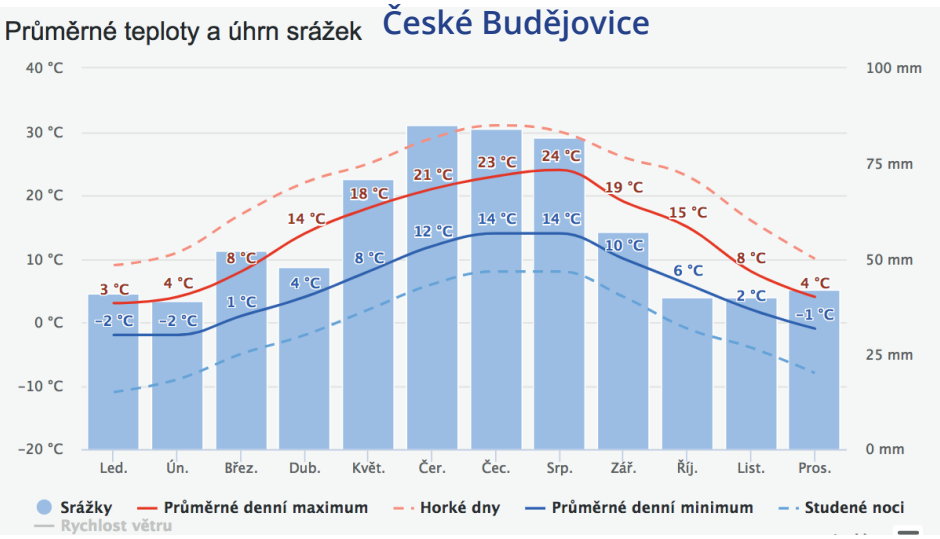
Data kolem nás

Je nám jasné, že z pojmu data ti v dnešní době může jít hlava kolem. To jsou data v mobilu, data v tabulkách a pak najednou někdo přijde a mluví o big data, tedy nějakých velkých datech. Čert aby se v tom vyznal. Proto ti přinášíme tento tutoriál, kde se ti pokusíme v pojmu data udělat aspoň trochu pořádek.

Co jsou data?

Data jsou údaje, kterými můžeme popsat vlastnosti pozorovaných objektů. Získáme je tak, že je zapíšeme, změříme a nebo vypozorujeme. Jinak řečeno – dokážeme tyto vjemy zachytit svými smysly. Data popisují objekty a události z reálného světa. Tak, a jak si teď vysvětlit předchozí věty?

Začneme příkladem. Určitě znáš teploměr a možná jej máš doma na okně. Teploměr je jedním z mnoha nástrojů, pomocí něhož můžeš data získat – neustále poskytuje informace o teplotě. Pokud si budeš pravidelně zaznamenávat naměřenou teplotu, vznikne ti tvá vlastní databáze. S ní můžeš poté pracovat a vypočítat, kolik byla průměrná teplota v sedm ráno za předchodí měsíc nebo zjistit, jaká byla nejvyšší a nejnižší teplota. Dokonce si vše můžeš překreslit do grafu a zkoumat, jak se jednotlivé dny lišily.



Zdroj: https://www.meteoblue.com/cs/

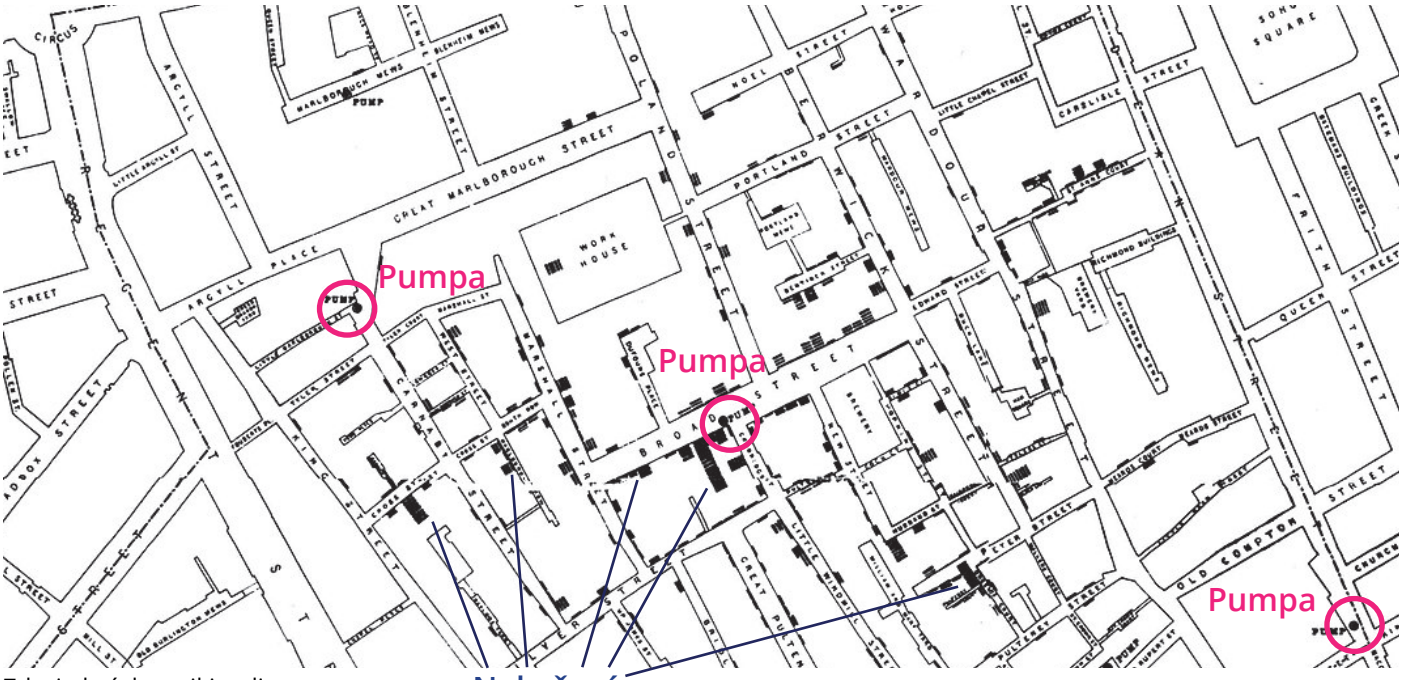
To, že si zapisuješ naměřenou teplotu, lze nazvat sběrem dat. A data se dají sbírat o lecčems. Třeba o tvých nákupech v Minecraftu. U každého nákupu totiž tvůrci hry vidí kdo si co koupil, v jakém množství a kolik utratil. Když si tedy pořídíš nový skin, jsou tyto informace uloženy v nějaké databázi (tabulce). Ta poté pomáhá výrobcům hru upravovat, aby byla pro hráče atraktivnější.

Jak je tento pojem starý?

Mohlo by se ti zdát, že pojem data a práce s nimi se zde objevuje teprve v posledních letech. Opak je ale pravdou. Možná se sice o termínu data mnohem více mluví až teď, ale s daty se pracuje již od druhé poloviny devatenáctého století. Přenesme se proto na chvilku do Londýna do roku 1832.

Město samotné rostlo a lidí přibývalo. Domy ještě neměly vlastní vodovodní potrubí a pro vodu se chodívalo k pumpám. Kanalizace tehdy také ještě nebyla dokonale vyřešena. Veškeré odpadní vody putovaly do žump, odkud mohly lehce prosáknout také ke zdrojům pitné vody. Jistě si umíš představit, co udělá voda z kanálu nebo žumpy s pitnou vodou… Nehledě na to, že lidé vylévali odpady také do řek. Řeka Temže tehdy určitě nevoněla a neměla ani zdánlivě modrou barvu.

V létě roku 1832 se v Londýně objevila cholera. To je nemoc, během které lidé hodně zvrací a mají průjem. Nebezpečná je v tom, že pacienti nestíhají dostatečně pít, a tak umírají na dehydrataci, tedy nedostatek vody. Lékař John Snow měl podezření, že by se tato nemoc mohla šířit vodou. Proto když se v padesátých letech devatenáctého století tato nemoc do Londýna vrátila, mohl John Snow své podezření potvrdit. Zaznamenával postupně počet nakažených a sledoval, z jaké pumpy berou vodu. Víš, jak se mu podařilo dokázat, že se nemoc opravdu šíří skrz vodu (tedy veřejné pumpy)? Jedněmi z nenakažených lidí byli zaměstnanci pivovaru. Přestože měli pumpu s (již znečištěnou) vodou nablízku, nepoužívali ji, ale pili vlastní pivo.



Zdroj obrázku: wikipedia.com

Kde všude dnes potkáš datovou analytiku

Odpověď je jednoduchá – skoro všude. Na příklad v obchodech a na poštách, kde často nebývají otevřené všechny pokladny či přepážky v dobu, kdy chodí málo lidí. Více pokladen pak otevírá třeba v dobu, kdy lidé končí v práci a míří na nákup. Nebo na YouTube. Stává se ti, že když zhlédneš jedno video, YouTube ti doporučí další, podobné? Může být na stejné téma nebo od stejného autora. To se také děje za pomoci datové analytike. Dále pak nesmíme opomenout e-shopy, které ti často doporučují dárky podle toho, co vyhledáváš. Takže pokud se díváš po nových slunečních brýlích, lehce se ti stane, že ti sama stránka nabídne další produkty, které by se ti mohly líbit (třeba kšiltovku).

Datové modelování

Datové modelování si nejlépe vysvětlíme na Minecraftu. Každý hráč má ve hře svůj účet – svůj profil. Hra samotná ví o každém hráči stejné věci – určitá data. A sice jak se hráč jmenuje, jakou má e-maiovou adresu, jak si vede jeho postava. Tedy kolik má předmětů v inventáři (a jakých), jaké má zakoupené skiny,... Aby si hra byla schopna tohle zapamatovat (a my nehráli pořád dokola od začátku), má v sobě zabudovaný nějaký systém tabulek, tedy nějakou databázi. Tato databáze v sobě udržuje informace během celého průběhu hry o tom, který hráč má jaký předmět. Pojďme se podívat, jak to může v tabulkách vypadat.

Hráč

ID	Jméno	E-mail
1	HonzaMinecraftak	honza.minecraftak@outlook.cz
2	Kuba_001	kuba_minecraft@email.cz

Hráč je tabulka, která nám udává informace o hráčích. Každý má nějaké identifikační číslo (ID), své jméno a e-mailovou adresu. Proč tolik údajů? Jména hráčů se mohou opakovat, nejsou tedy unikátní, proto je potřeba, aby bylo ve hře opravdu jasné, komu patří který účet. Proto používáme ID. Nemohou existovat dva hráči, kteří by měli stejné číslo.
Je jasné, že čím později se ke hře přidáš, tím vyšší ID budeš mít.

Předmět

ID	Jméno	Popis	Materiál
1	Meč	Pro útok	Železo
2	Štít	Pro útok	Dřevo

V tabulce *Předmět* zase vidíme, jaké věci (předměty) mají hráči k dispozici. Každý předmět má své unikátní identifikační číslo, a to proto, že některé předměty jsou si podobné. Hra si musí být jista, že žádné dva předměty nezamění. Dále je v tabulce jméno předmětu, popist, k čemu slouží a také materiál, z něhož je věc vyrobena.

Inventář

ID hráče	ID předmětu	Akce	Datum a čas akce
1	1	Přidání do inventáře	17.03.2018 15:02
2	2	Přidání do inventáře	17.03.2018 15:06
1	1	Odebrání z inventáře	17.03.2018 15:09
2	1	Odebrání z inventáře	17.03.2018 15:10
2	2	Přidání do inventáře	17.03.2018 15:12
1	2	Přidání do inventáře	17.03.2018 15:13

A co vidíme v tabulce *Inventář*? Představ si, že Honza s Kubou si spolu chtějí zahrát, takže se přihlásí do hry. Každého hráče musíme mít s unikátním identifikačním číslem (ID), protože spolu mohou hrát i dva Jirkové ze zřídy, tak ať je jasné, kdo je kdo. Toto unikátní ID (tedy takové, které jiný hráč nemá) se v datové analytice nazyvá primární klíč. Je to něco, co jednoznačně identifikuje (rozpozná) daného hráče.
Ve hře jsou dva předměty, jeden meč a jeden štít. Jakmile začne hra, Honza sebere meč. Kuba si zase vezme štít. V tabulce *Inventář* jednotlivé hráče rozlišujeme pomocí ID hráče, což jsou primární klíče v tabulce *Hráč*. Tabulka *Inventář* si ale tyto klíče vypůjčuje z jiné tabulky. Jsou tu proto nazývány jako cizí klíč.

Tip! Zjisti, kdo jako poslední odkládá štít a kdo jej drží nakonec? Který hráč končí tah s mečem v ruce?

Dimenze a metriky

Ukládání dat v Minecraftu bychom tedy měli. Pro hru samotnou je tento způsob dobrý, co je složitější, je v tuto chvíli pokládání otázek. Určitě se chceme během hry pravidelně ptát na otázky typu „Který hráč má nejvíce předmětů?“ či „Který předmět má nejvíce hráčů v inventáři?“. No a upřímně, pro tyto zvídavé otázky je výše popsaný způsob uložení dat zbytečně komplikovaný.

Abychom si mohli odpovídat na otázky, vznikl jiný způsob uložení dat, který používá jednoduchý princip dimenzí a metrik. I když to zatím zní složitě, není.

Uveďme si příklad: počet hráčů Minecraft na světě.
Když se mluví o počtu, je jasné, že půjde o nějaké číslo, třeba 144 000 000 (sto čtyřicet čtyři miliónů). Toto je metrika. Metrika je jednoduše číslo, se kterým se dají dělat operace jako zjisťování nejmenšího čísla, největšího čísla, průměry, mediány a podobně. Počet hráčů bychom tedy měli.
A teď pojďme k „na světě.“ Od tohoto výrazu můžeme chtít trochu více – třeba „v jednotlivých zemích,“ což znamená, že bychom číslo 144 000 000 rozdělili mezi státy, kde se Minecraft hraje. V USA by to mohlo být 50 miliónů, v Německu pak 3 milióny a v České republice pak půl milionu hráčů, … Až bychom vyjmenovali všechny země a jejich počty hráčů, tak bychom je pro kontrolu sečetli. A vyšlo by nám opět původní číslo (144 000 000). Tohle je jedna dimenze.

Dvě dimenze pak fungují tak, že existují všechny kombinace všech dimenzí. Uf, pojďme si to ukázat na příkladu, než se v tom všem ztratíme.

Máme naše hráče (těch 144 000 000), které jsme rozdělili do zemí. My bychom k tomu ale také rádi věděli, jaká zařízení kdo používá (jestli Xbox nebo počítač). A zjistíme následující… (tabulka je neúplná a slouží pro demonstraci).

Země	Zařízení	Počet hráčů
USA	PC	29 450 000
USA	Xbox	20 550 000
Německo	PC	2 200 000
Německo	Xbox	800 000
Česká republika	PC	300 000
Česká republika	Xbox	200 000

Takže tady vidíme, že pro každou zemi musím uvést číslo pro každé zařízeí. Máme tedy 3 země a 2 zařízení: 2 x 3 = 6 řádků s číslem, neboli metrikou. A tento formát už je tedy lepší pro dotazování – můžeme se ptát, ktaré zařízení je nejob-

líbenější, ve které zemi na světě je nejvíce hráčů Xboxu či kolik je v každé temi průměrně hráčů hrajících na počítači.

A teď se vraťme zpět k našemu příkladu s hráči a jejich předměty. Po takzvané "transformaci" dat (data jenom jinak vyjádříme, ale neměníme jejich význam; je to jako když zmrazíme vodu na led – je to pořád voda, ale mé jiné skupenství – to samé uděláme s daty), kterou jsme provedli a dostali se z našeho původního modelu dobrého pro hru (dobře se ukládají a mění hráči a předměty) až k modelu dobrému pro dotazování datovými analytiky. A ten má už jenom jednu tabulku, zato více sloupců. Může vypada třeba takto (pro lepší příklad si přimyslíme více dat, než jsme měli nahore).

Hráč	**Předmět**	**Hodina hry**	**Počet přidání do inventáře**	**Počet odebrání z inventáře**
Honza	Meč	15.	1	1
Honza	Štít	15.	1	0
Kuba	Meč	15.	1	0
Kuba	Štít	15.	1	1
Honza	Meč	16.	2	0
Honza	Štít	16.	1	0
Kuba	Meč	16.	0	1
Kuba	Štít	16.	0	1

Ttohle je nasše tabulka, která je super na odpovídání na otázky. Takže třeba - který předmět byl nejvícekrát přidán do inventáře? Stačí, když se podíváš na tabulku a sečteš všechna čísla *Počet přidání do inventáře* pro jednotlivé předměty. Získáš tak informace jako Meč = 4x, Štít = 3x. A na to samé se můžeš zeptat u hráčů – který z nich nejvícekrát posbíral nějaký předmět? Sečteš si čísla pro jednotlivé hráče a je to. Takhle mmůžeš vymyslet spoustu otázek typu „Ve které hodině hráči nejvíce odhazovali předměty?" „Který hráč má aktuálně nejvíce předmětů?" U té poslední otázky to chce trošku obezřetnosti, protože musíš spočítat, *Počet přidání do inventáře* mínus *Počet odebrání z inventáře*.

Datového modelování stačí, pojďme na vizualizace!

Datové vizualizace

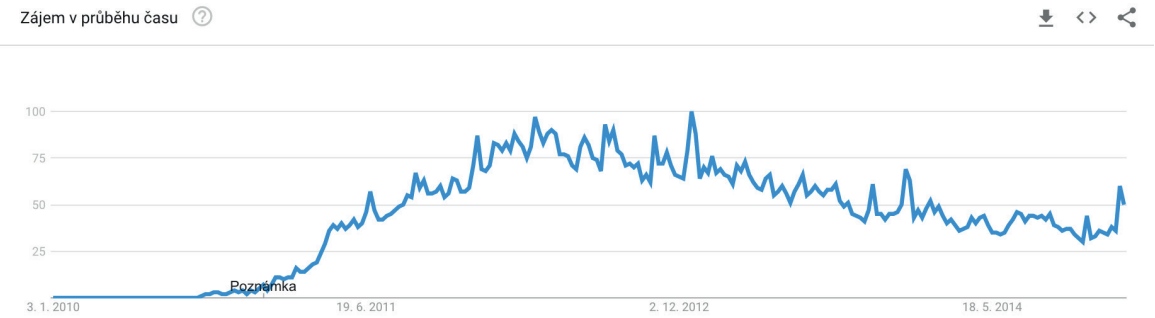
Datoví analytici ani lidé obecně nejsou moc dobří v tom, aby rozuměli datům jenom jako číslům v tabulkách. Je těžké si pro nás představit číst v nich různé odpovědi na otázky – je to totiž náročné na přemýšlení, na soustředění a je to strašná nuda.
Říká se, že obrazek vydá za tisíc slov. Co to znamená? Stačí jeden obrázek a mnohd ypochopíme více, než kdybychom přečetli celou stránku textu. Je tomu tak i v analytice, jen ten obrázek nevydáaza tisíc slov, ale milion čísel.

Pamatuješ na Jona Snowa? Tak jeho bádání začalo jako zápisky v sešitě. Aby ale objevil souvislost vodních pump s nemocí, musel si všechno rozkreslit a namalovat. To by jen ze svých zápisků a tabulek nepoznal. I proto si rádi v datové analytice data nejdříve zpracujeme a připravíme, ale ihned potom vizualizujeme – neboli nakreslíme.

Pojďme si dát příklad. Co ti řekne následující tabulka? Jak se vyvíjí hledanost Minecraftu ve vyhledávači v ČR?

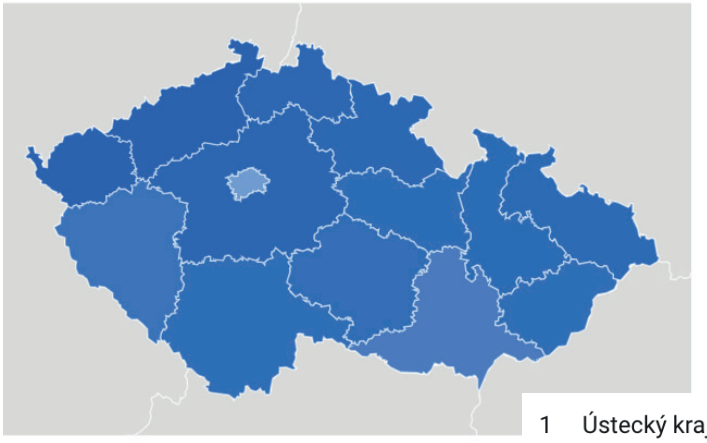
2010-01-03,0	2010-05-23,0	2010-10-10,3
2010-01-10,0	2010-05-30,0	2010-10-17,2
2010-01-17,0	2010-06-06,0	2010-10-24,2
2010-01-24,0	2010-06-13,0	2010-10-31,3
2010-01-31,< 1	2010-06-20,0	2010-11-07,4
2010-02-07,0	2010-06-27,0	2010-11-14,3
2010-02-14,< 1	2010-07-04,0	2010-11-21,4
2010-02-21,0	2010-07-11,< 1	2010-11-28,2
2010-02-28,0	2010-07-18,< 1	2010-12-05,4
2010-03-07,0	2010-07-25,< 1	2010-12-12,3
2010-03-14,0	2010-08-01,0	2010-12-19,5
2010-03-21,< 1	2010-08-08,< 1	2010-12-26,7
2010-03-28,0	2010-08-15,0	2011-01-02,4
2010-04-04,0	2010-08-22,< 1	2011-01-09,7
2010-04-11,< 1	2010-08-29,< 1	2011-01-16,11
2010-04-18,0	2010-09-05,< 1	2011-01-23,11
2010-04-25,< 1	2010-09-12,1	2011-01-30,10
2010-05-02,0	2010-09-19,2	2011-02-06,11
2010-05-09,< 1	2010-09-26,2	2011-02-13,11
2010-05-16,0	2010-10-03,3	...

No... nám tohle nic neřekne. Co když si zkusíme dát čísla do grafu?



Tohle je mnohem lepší. Dá se najednou jednoduše poznat, že zajem o pojem Minecraft byl nejvyšší v červenci roku 2012, od té doby je mírně na ústupu. Vysvětlení může být několik – kolem roku 2012 již málokdo potřeboval zjišťovat, co je ten záhadný Minecraft. Hra se mohla dostat do povědomí více lidí, čímž mizel důvod toto klíčové slovo vyhledávat.

Grafy ale nemusí být pouze nudné čáry zleva doprava. Dají se různě kombinovat a existuje jich nepřeberné množství. Dokonce i takové, které se dají zakreslit na mapu. Pojďme se tedy podívat, ve kterém kraji ČR je Minecraft nejoblíbenější:



1	Ústecký kraj	100
2	Karlovarský kraj	98
3	Středočeský kraj	90
4	Liberecký kraj	88
5	Královéhradecký kraj	86

V námi sledovaných letech to byl Ústecký kraj s Karlovarským krajem v patách.

Kam dál?

A je to... Ukázali jsme si, co jsou to data, jaké mají formy a co se s daty dá dělat. Pokud tě bude datová analytika zajímat dál, máš na výběr z několika různých zaměření.

1. Můžeš jít cestou datového modelování a datové vizualizace, které jsme si v tomto tutoriálu ukázali. Budeš tak pomáhat ostatním chápat data kolem nich (a vytvářet ony obrázky, jež vydají za více než milion řádků v tabulkách). Tady se budeš muset naučit, jak ovládat vizualizační nástroj (neboj, stačí ti i Excel, ale existují i další nástroje, které jsou stvořené pro pohodlnou práci s vizualizací, třeba PowerBI).

2. Možná tě zaujme datové inženýrství. To se stará o to, aby byla data řádně posbírána (například z teploměru nebo z aplikace) a uložena tak, aby se jich datový analytik mohl dotazovat (tak, jak jsme se dnes dotazovali my, ale my jsme měli jednoduché tabulky, samozřejmě). Tato disciplína je více o programování a porozumění databázím.

3. Prozkoumej, co se skrývá pod pojmem datová věda. Ta je blíže matematice. Na základě již posbíraných dat se snaží předpovídat budoucnost. To datovým vědcům někdy jde, někdy zase trochu méně. Nicméně jde o důležitou část pro všechny firmy a instituce.

Ať už se vydáš kamkoli, věz, že svět dat je velmi rozmanitý a je se zde vždy co učit. Tak s chutí do toho!

Tento materiál vznikl v rámci projektu Akademie programování, na němž spolupracují organizace Czechitas s firmou Microsoft.

Našli jste v textu nesrovnalosti? Veškeré dotazy, náměty a komentáře prosím směřujte na **paja@czechitas.cz**, který vede k Pavle Randákové. Pavla působí v organizaci Czechitas na pozici Youth Education specialist.