

Ukládání rozsáhlých dat v NoSQL databázích

František Koleček, xkolec08@stud.fit.vut.cz

Tomáš Moravčík, xmorav41@stud.fit.vut.cz

David Sladký, xsladk07@stud.fit.vut.cz

zima 2022

Obsah

| | | |
|-----------|---|----------|
| I | Analýza zdrojových dat a návrh jejich uložení v NoSQL databázi | 1 |
| 1 | Analýza zdrojových dat | 2 |
| 2 | Návrh způsobu uložení dat | 3 |
| 3 | Zvolená NoSQL databáze | 4 |
| II | Návrh, implemetace a použití aplikace | 5 |
| 4 | Návrh aplikace | 6 |
| 5 | Způsob použití | 7 |
| 6 | Experimenty | 8 |

Část I

Analýza zdrojových dat a návrh jejich uložení v NoSQL databázi

Kapitola 1

Analýza zdrojových dat

Cíl: Zjistit a popsat, jaké datové sady budou pro úlohu v projektu potřeba a jaké jsou jejich vlastnosti tak, abychom mohli zvolit vhodný formát a způsob uložení v NoSQL databázi.

Obsah: Pro významné použité datové soubory určit nad rámec jejich veřejně přístupné dokumentaci alespoň následující: odkud a jak ho lze získat, jak často a jakým způsobem je tam aktualizován, jaké jsou (jen významné) části jeho struktury (pokud je strukturovaný, tj. elementy/sloupce a jejich datové typy, povinnost/volitelnost, opakování, atp.; jen dále významné části, pokud pro zbytek lze udělat odkaz do veřejné dokumentace ke zdroji, existuje-li taková), jaký je identifikátor položek ve zdroji (pokud existuje) a v jakém kontextu/rozsahu a časovém rozmezí jsou jeho hodnoty jedinečné, jaké jsou případné odkazy z položek zdroje na položky dalších zdrojů (pokud tam takové jsou), jaké jsou možnosti seskupení položek ze zdroje na základě jejich hodnot (např. zaměstnance můžeme seskupt dle oddělení), případné speciální rysy dat ve zdroji (např. temporální či multimediální složky, geografická data, atp.), další poznámky důležité pro následný návrh (např. velikost zdroje, jeho chybovost, atp.).

Prostředky: Stručný strukturovaný text, odrážky, případně tabulky; případně s odkazy na konkrétní místa do veřejné dokumentace zdroje na Internetu.

Fáze projektu: Hned na začátku po seznámení se zdroji dat.

Kapitola 2

Návrh způsobu uložení dat

Cíl: Po posouzení vlastností dat (z předchozí analýzy) a očekávaných dotazů (ze zadání) navrhnout vhodný způsob uložení dat do NoSQL databáze. Způsob uložení musí být vhodný z hlediska způsobu nahrávání dat ze zdroje do databáze (a to i průběžného doplňování či aktualizace, bez smazání celé databáze) a z hlediska rychlosti dotazování dat v databázi z aplikace s využitím vlastností NoSQL (s využitím klíčů a škálovatelnosti/distribovanosti databáze). Data lze při nahrávání ze zdroje do databáze předzpracovávat, např. kombinovat či doplňovat, odvozovat pomocná data, předpočítávat agregace, atp. Takové předzpracování může trvat déle (kritérium vhodnosti při předzpracování v průběhu nahrávání není čas, ale vhodné využití obecných vlastností NoSQL, jako je sharding).

Obsah: Pro skupinu či každou podstatnou vlastnost dat z analýzy a dotaz ze zadání (pokud bude mít vliv na návrh) popsat, co znamená, jaký problém představuje, jaké je řešení, proč je zvolené řešení dobré a stručně jaké jsou případné alternativy.

Prostředky: Strukturovaný text (odstavce, sekce, odrážky, atd.), kde je popsán proces získání, předzpracování a uložení dat ze zdroje do databáze. Možno použít také pseudokód či diagramy popisující datové toky a použité struktury a vlastnosti NoSQL databází obecně. Každé návrhové rozhodnutí musí být řádně zdůvodněno (např. části se strukturou "dotaz/vlastnost", "problém", "řešení", "důvod", "alternativy").

Fáze projektu: Po analýze dat a analýze uživatelských požadavků na aplikaci, většinou souběžně s návrhem aplikace.

Kapitola 3

Zvolená NoSQL databáze

Cíl: Rozhrnout a zdůvodnit jaký druh NoSQL databáze je vhodný (zdůvodnění plyne částečně již z předchozího návrhu) a jaký konkrétní produkt NoSQL databáze bude použit.

Obsah: Určit typ databáze a konkrétní produkt NoSQL, vypsát jeho vlastnosti, které jsou pro toto řešení užitečné (a jiné než u jiných typů a produktů NoSQL) a zdůvodnit jejich vhodnost v kontextu předchozího návrhu.

Prostředky: Stručný volný text (až několik kratších odstavců) s případným vyznačením podstatných částí.

Fáze projektu: Zakončování návrhu a přechod k implementaci.

Část II

Návrh, implemetace a použití aplikace

Kapitola 4

Návrh aplikace

Cíl: Navrhnout hlavní části aplikace splňující požadavky zadání s důrazem na práci s na ni napojenou databází NoSQL či datovými zdroji (při jejich předzpracování a nahrávání do NoSQL databáze).

Obsah: Použité technologie (např. skriptovací jazyk, knihovny, atp.) a architektura (např. skript či sekvence skriptů pravidelně spouštěných v daných časových intervalech či v reakci na danou událost). Způsob technického řešení úloh ze zadání (jejich průběh v aplikaci) a konceptů z předchozího návrhu (struktury, algoritmy, toky dat, atp.).

Prostředky: Strukturovaný text (sekce, odstavce, odrážky, atp.), případně pseudokód či obrázky, doplňující technické detaily konceptů nastíněných v předchozím návrhu. Důraz je kladen na způsob realizace dotazů ze zadání.

Fáze projektu: Návrh aplikace a částečně po či souběžně s návrhem databáze.

Kapitola 5

Způsob použití

Cíl: Poskytnout stručnou dokumentaci pro zprovoznění databáze a aplikace.

Obsah: Stručně popsat, jak celé řešení zprovoznit, tj. nasadit databázi i aplikaci vč. způsobu volání aplikace (příkazový řádek, parametry) pro úlohy předzpracování a nahrání dat ze zdroje do databáze a pro úlohy hledání nad databází tak, jak byly definovány v zadání.

Prostředky: Stručný text obsahující návod (popis) s ukázkami způsobu volání aplikace (např. pro skripty by to byl kód příkazového řádku).

Fáze projektu: Dokončování implementace, chystání dokumentace pro předání výsledného systému zákazníkovi.

Kapitola 6

Experimenty

Cíl: Změřit, jak aplikace a databáze fungují v praxi.

Obsah: Popis výchozí konfigurace aplikace a nasazení databáze stroje, kde budou experimenty probíhat (HW a SW). Popis experimentů typicky představující nahrání dat ze zdroje do databáze či dotazy ze zedání s výslednými časy jejich provedení. Případné poznámky k výsledkům experimentů.

Prostředky: Strukturvaný text, případně tabulka či graf s doprovodným textem.

Fáze projektu: Testování řešení před předáním výsledného systému zákazníkovi.