Technologie pro velká data (B0M33BDT) – domácí úkol

Data

Adresní místa registru RÚIAN k 31. 5. 2020. Na Metacentru k dispozici v lokálním FS (tj. ne na HDFS): */home/pascepet/fel_bigdata/data/ruian.zip*

Soubor .zip obsahuje tři adresáře, z nichž budete pracovat jen s adresářem *adresni_mista*. V tomto adresáři je něco přes 6 tisíc CSV souborů, každý obsahuje adresní místa pro jednu obec ČR. Kódování CSV souborů je WINDOWS-1250.

Zadání

1. Z dat všech adresních míst vytvořte na Hadoopu managed tabulku Hive. Formát a kompresi zvolte podle svého uvážení, názvy jednotlivých sloupců si také zvolte sami (viz komentáře níže). Podmínky jsou jen tyto: v tabulce musejí být jen řádky za adresní místa, žádné řádky s původními názvy sloupců; data budou uložena v kódování UTF-8.
2. Zjistěte pět nejčastějších názvů ulic (s největším počtem adresních míst), prázdné názvy ulic se samozřejmě nepočítají.
3. Ověřte, zda u adresních míst, která mají Typ SO „č.p.“, platí pro první číslice domovního čísla Benfordův zákon (https://cs.wikipedia.org/wiki/Benford%C5%AFv_z%C3%A1kon).
4. Vypočítejte souřadnice těžiště ČR podle všech adresních míst a těžiště pro každou obec zvlášť (obec je definovaná kódem obce). Poté najděte pět obcí, jejichž těžiště jsou nejblíž těžišti celé ČR, a pět obcí, jejichž těžiště jsou nejdál od těžiště celé ČR. Vzdálenosti těžišť obcí od těžiště ČR zjistěte v kilometrech. Viz komentář níže ohledně souřadnic S-JTSK.
5. Zjistěte, které adresní místo je v ČR nejzápadnější. Podobně zjistěte adresní místa nejsevernější, nejvýchodnější a nejjižnější. K tomu budete potřebovat převést souřadnice S-JTSK na GPS, viz komentář níže.

Technologie

Výpočty jsou bez problémů proveditelné na výpočetním clusteru Metacentrum. Očekává se, že použijete Linux, Hadoop, HDFS, Hive, případně Spark. Volba konkrétního nástroje v každém zadání (zejména 2–5) je na vás.

Očekávané výstupy

1. Existence managed tabulky Hive podle zadání.
2. Textový soubor (nebo více souborů) se zdrojovým kódem skriptů k jednotlivým bodům zadání (Hive, Spark). Název databáze a tabulky z bodu 1 musí odpovídat názvům ve zdrojovém kódu. Není nutno uvádět linuxové a HDFS příkazy.
3. Odpovědi pro body 2–5 zadání. Je možné tyto odpovědi vložit do textového souboru se zdrojovými skripty nebo poslat samostatně.

Zdrojové skripty a odpovědi zašlete e-mailem na adresu jan.hucin@profinit.eu.

Hodnocení

Za domácí úkol je možné získat max. 10 bodů. Každá z pěti částí zadání se hodnotí max. dvěma body s granularitou půl bodu. Podstatné jsou: vhodnost zvoleného postupu, správnost zápisu a funkčnost (správnost odpovědí).

Termín

Půlnoc z 31. 12. 2020 na 1. 1. 2021.

Komentáře

- V datech jsou uvedeny sloupce v tomto pořadí:
  kód (id) adresního místa, kód obce, název obce, kód městské části nebo obvodu, název městské části nebo obvodu, kód starého městského obvodu (jen Praha), název starého městského obvodu, kód části obce, název části obce, kód ulice, název ulice, typ domovního čísla (č.p. nebo č.ev.), domovní číslo, číslo orientační („modré“), přidaný znak k orientačnímu číslu (např. u adresy „Valdštejnská 6a“), poštovní směrovací číslo, Y a X souřadnice systému S-JTSK, datum platnosti záznamu.
- Existují různé obce, které mají stejný název (ale samozřejmě různé kódy obcí). Proto je potřeba rozlišovat obce (v bodu 4 zadání) podle kódu, ne jen podle názvu.
- Ze zadání plyne, že nepotřebujete všechny sloupce, tj. např. se nemusíte trápit s uložením nestandardního datumu platnosti jako *date*, můžete ho nechat jako *string* – apod.
- Konzola Beeline má bug, kdy znaky obsahující háčky nebo čárky zobrazuje jako otazníky. Také je problém přes konzolu Beeline zadat SQL dotaz obsahující takovéto znaky (např. jako podmínku). Je nutné a naštěstí ne složité to nějak obejít.
- Souřadnice adresních míst jsou zadány v málo známém, ale velmi praktickém systému S-JTSK (https://cs.wikipedia.org/wiki/Syst%C3%A9m_jednotn%C3%A9_trigonometrick%C3%A9_s%C3%ADt%C4%9B_katastr%C3%A1ln%C3%AD). Jednotky systému jsou metry, je to pravoúhlá síť, takže vzdálenost mezi dvěma body se počítá velmi snadno. Pouze osy systému nejsou orientované severojižně a západovýchodně. Pro bod 4 zadání to ale nevadí, tam světové strany nehrají žádnou roli.
- Pro bod 5 zadání je nutné provést přepočet S-JTSK na GPS. Funkci pro Javascript je možné vyextrahovat ze zdrojového kódu webové stránky http://martin.hinner.info/geo/, je též uložená v samostatném souboru u zadání na Courseware. Pro použití v Hive nebo Sparku je pochopitelně potřeba funkci přepsat do Pythonu, Javy nebo něčeho podobně použitelného. Je to ovšem snadné, výpočet je sice dlouhý, ale přímočarý.
- Těžiště z adresních míst se spočítá přímočaře jako průměr všech X souřadnic a průměr všech Y souřadnic. (Je to vlastně výpočet hmotného středu soustavy konečného počtu hmotných bodů – adresních míst, kde každé adresní místo má stejnou váhu.)