Obrázok, na ktorom je text, hodiny

Automaticky generovaný popisObrázok, na ktorom je logo

Automaticky generovaný popisEKONOMICKÁ UNIVERZITA

Fakulta hospodárskej informatiky

Katedra aplikovanej informatiky

Analýza neštruktúrovaného a semištruktúrovaného zdroja pomocou Sparku

**Predmet:** Big Data

**Meno študentov:** Martin Jankech, Patrik Hajdučík

**Stupeň/ročník:** druhý/2.ročník

**Školský rok:** 2022/2023

**Meno prednášajúceho:** Ing. Jaroslav Kultán PhD.

**Meno cvičiaceho:** SCHMIDT, Peter, Ing. Mgr., PhD.

Obsah

[Úvod 3](#_Toc131700291)

[Inštalácia 4](#_Toc131700292)

[Vybrané dáta 6](#_Toc131700293)

[Selecty 7](#_Toc131700294)

[Záver 9](#_Toc131700295)

[Zdroje 10](#_Toc131700296)

# Úvod

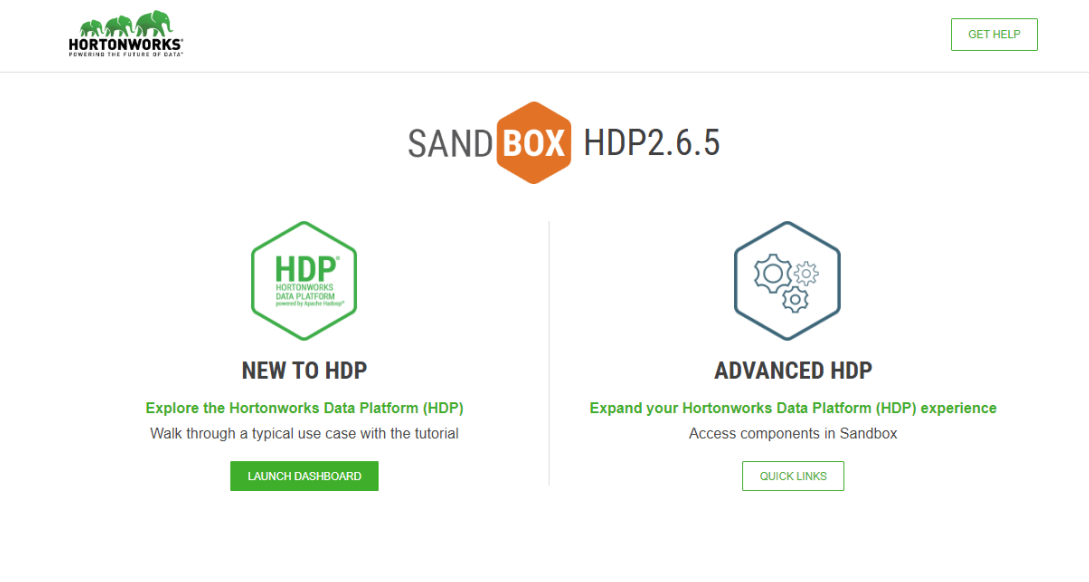
V dnešnej dobe sa stáva stále bežnejšie, že organizácie potrebujú analyzovať a spracovávať obrovské množstvá dát. HIVE je nástroj, ktorý umožňuje spracovanie dát uložených v distribuovanom súborovom systéme Hadoop pomocou SQL-like dotazovacieho jazyka. Pre správne používanie HIVE a efektívne spracovanie dát je dôležité mať nástroj na správu Hadoop klastra. Jedným z najznámejších nástrojov na správu Hadoop klastra je Ambari.

# Inštalácia

Nainštalovanie Ambari je pomerne jednoduché a môžete to urobiť pomocou nasledujúcich krokov:

Vyberte si distribúciu Hadoopu, ktorú chcete použiť. Ambari podporuje rôzne distribúcie Hadoopu, ako napríklad Hortonworks, Cloudera alebo Apache.

Na spustenie ambari je potrebné mať nainštalovaný hypervízor, napríklad v našom prípade vol použitý Oracle VirtualBox ktorý umožní spustiť virtuálny stroj s Hortonworks Sandbox HDP 2.6.5.



Meno a Heslo je maria\_dev.

Po spustení je potrebné pozapínať všetky služby. V prípade erorov viete do systému pristúpiť pomocou Putty a príslušné erory vyriešiť.

Následne si v pravom hornom rohu rozbalíme menu a vyberieme Hive View.

Obrázok, na ktorom je text

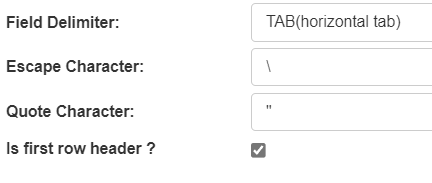
Automaticky generovaný popis

V zobrazenom Query editore si vytvoríme databázu a po kliknutí execute prejdeme na panel Upload Table.

Obrázok, na ktorom je text

Automaticky generovaný popis

Následne nahráme súbory do príslušnej databázy vo forme tabuliek. Nezabúdame na určenie delimitera a v prípade ak zdrojové údaje importujeme aj s atribútmi zaklikneme túto možnosť aj v nastaveniach.



Po importnutí tabuliek sa vrátime do sekcie Query refreshneme našu db a vkladáme Selecty.

# Vybrané dáta

V našej práci sme si vybrali a stiahli dataset zo stránky IMDb priloženej v zdrojovej časti práce. IMDb (Internet Movie Database) je online databáza filmov, televíznych programov, hereckých výkonov, režisérskych prác a iných informácií o filmovom priemysle. Stránka sa zaoberá poskytovaním informácií o rôznych filmoch a televíznych programoch.

Obrázok, na ktorom je text

Automaticky generovaný popisVybrané dáta pozostávali z dvoch databáz, ktoré obsahovali nasledovné atribúty:

Obrázok, na ktorom je text

Automaticky generovaný popis

Obrázok, na ktorom je text

Automaticky generovaný popis

# Selecty

Tento dotaz spojí obe tabuľky podľa atribútu "tconst" a potom zoskupí výsledky podľa "titleType". Funkcia "AVG" vypočíta priemerné hodnotenie pre každý žáner tak, že si zoberie hodnoty z tabuľky ratings (averageRating), ktoré predstavujú priemernú hodnotu hodnotenia, akú dostal daný produkt a následne vypočíta priemernú hodnotu všetkých už spomínaných priemerných hodnôt, spadajúcich do danej kategórie. Následne výsledky zoradí zostupne podľa výšky hodnotenia.

Obrázok, na ktorom je tabuľka

Automaticky generovaný popis

Obrázok, na ktorom je stôl

Automaticky generovaný popis

Vzhľadom na to, že každý hodnotený produkt má rovnakú váhu, pri výpočte našej finálnej priemernej hodnoty sme pridali obmedzenie pre produkty, ktoré mali menej ako 100 hlasov. Takéto produkty neboli v našich výpočtoch zohľadnené. Taktiež sme pridali informáciu o počte produktov v danej kategórií, ktoré túto podmienku spĺňali a teda predstavujú konkrétne číslo produktov, ktoré boli zohľadnené pri výpočte priemernej hodnoty žánru.

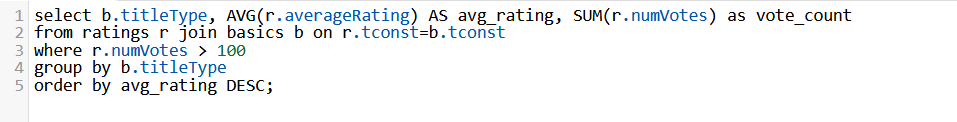
Obrázok, na ktorom je text

Automaticky generovaný popis

Obrázok, na ktorom je stôl

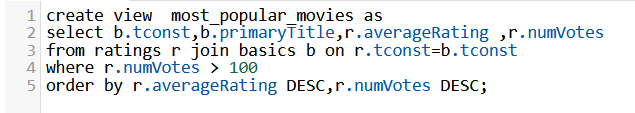
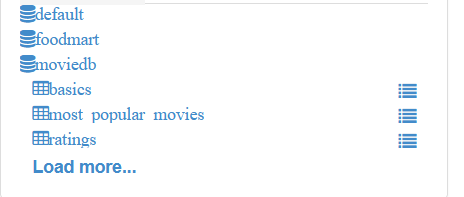
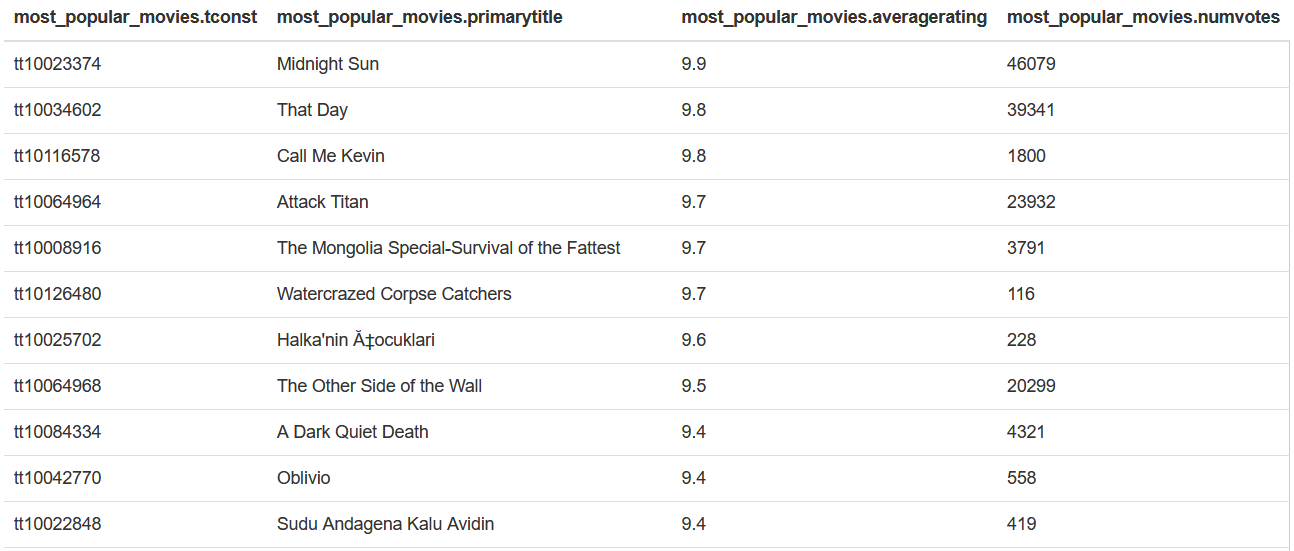
Automaticky generovaný popis

Select sme upravili a namiesto count sme použili sum čím sme získali celkový počet hlasov pre každu kategóriu. (tu bol použitý zredukovaný dataset (filmy od 2016-cca 35 000 filmov) preto vyšli aj menšie čísla).



# 

Vytvorili sme si aj nový pohľad, v ktorom sme si vybrali najpopulárnejšie filmy. Filmy sú zoradené najprv podľa priemerného hodnotenia a potom podľa počtu hlasov.



# Záver

HIVE a Ambari sú dva nástroje, ktoré sa dnes používajú na spracovanie a správu veľkých dátových množstiev. HIVE umožňuje spracovávanie dát uložených v Hadoop distribuovanom súborovom systéme pomocou SQL-like dotazovacieho jazyka. Ambari na druhej strane umožňuje jednoduchú a efektívnu správu Hadoop klastra. Použitie týchto nástrojov môže organizáciam umožniť spracovanie obrovského množstva dát rýchlo a efektívne.

# 

# Zdroje

[https://datasets.imdbws.com](https://datasets.imdbws.com/)

https://www.imdb.com/interfaces/?fbclid=IwAR2UMe3\_WL9YpoVLu4GuDY5q7kXai9NIUm9GNp3lFGKatRrPL7GYP8yk\_PM