



EKONOMICKÁ UNIVERZITA
Fakulta hospodárskej informatiky
Katedra aplikovanej informatiky



HIVE - ZÁKLADY PRÁCE V AMBARI

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
Inštalácia.....	4
Vybrané dáta	6
Selecty	7
Záver.....	9
Zdroje	12

Úvod

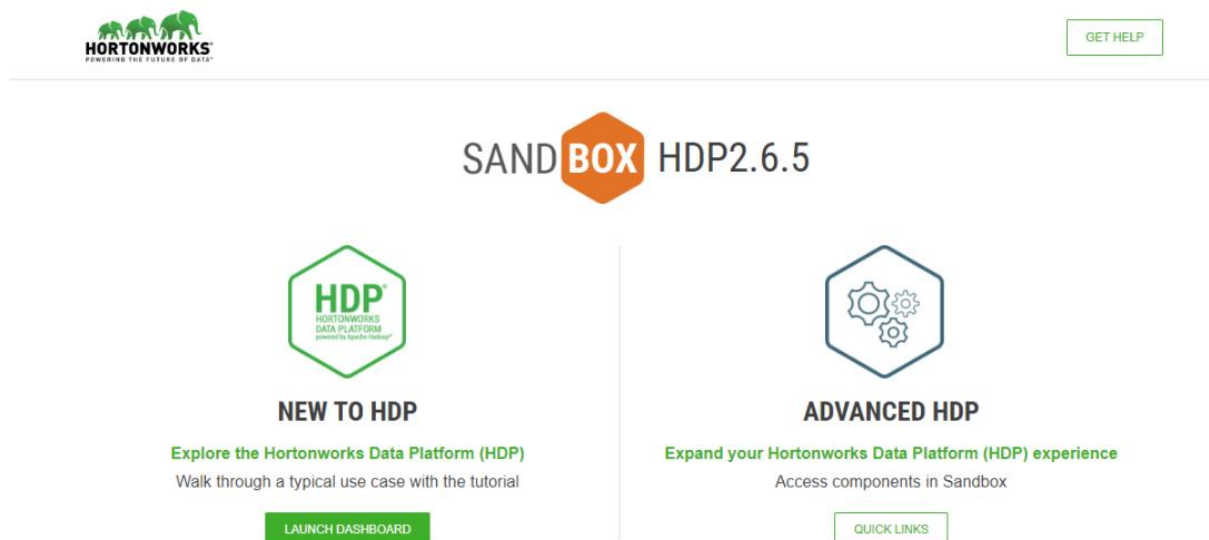
V dnešnej dobe sa stáva stále bežnejšie, že organizácie potrebujú analyzovať a spracovávať obrovské množstvá dát. HIVE je nástroj, ktorý umožňuje spracovanie dát uložených v distribuovanom súborovom systéme Hadoop pomocou SQL-like dotazovacieho jazyka. Pre správne používanie HIVE a efektívne spracovanie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

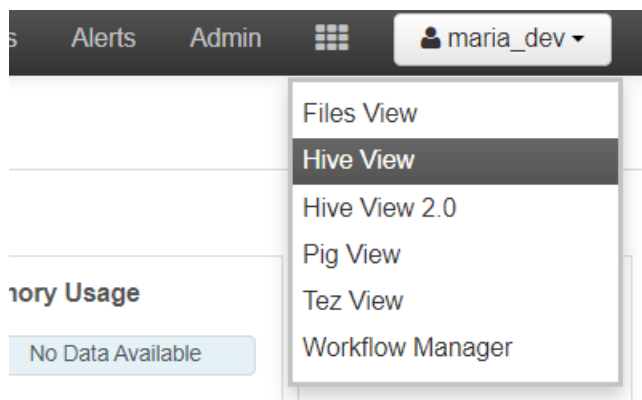
Na spustenie ambari je potrebné mať nainštalovaný hypervízor, napríklad v našom prípade vol použitý Oracle VirtualBox ktorý umožní spustiť virtuálny stroj s Hortonworks Sandbox HDP 2.6.5.



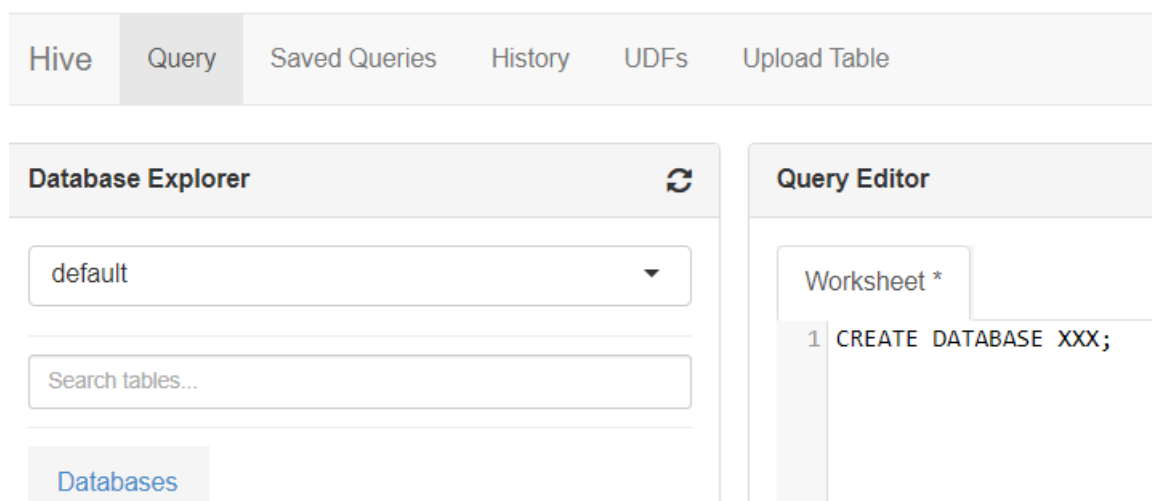
Meno a Heslo je maria_dev.

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

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



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



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.

Field Delimiter:	<input type="text" value="TAB(horizontal tab)"/>
Escape Character:	<input type="text" value="\"/>
Quote Character:	<input type="text" value=""/>
Is first row header ?	<input checked="" type="checkbox"/>

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.

IMDb data files available for download

Documentation for these data files can be found on <http://www.imdb.com/interfaces/>

[name.basics.tsv.gz](#)

[title.akas.tsv.gz](#)

[title.basics.tsv.gz](#)

[title.crew.tsv.gz](#)

[title.episode.tsv.gz](#)

[title.principals.tsv.gz](#)

[title.ratings.tsv.gz](#)

Vybrané dáta pozostávali z dvoch databáz, ktoré obsahovali nasledovné atribúty:

title.basics.tsv.gz - Contains the following information for titles:

- tconst (string) - alphanumeric unique identifier of the title
- titleType (string) - the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
- primaryTitle (string) - the more popular title / the title used by the filmmakers on promotional materials at the point of release
- originalTitle (string) - original title, in the original language
- isAdult (boolean) - 0: non-adult title; 1: adult title
- startYear (YYYY) - represents the release year of a title. In the case of TV Series, it is the series start year
- endYear (YYYY) - TV Series end year. 'N' for all other title types
- runtimeMinutes - primary runtime of the title, in minutes
- genres (string array) - includes up to three genres associated with the title

title.ratings.tsv.gz - Contains the IMDb rating and votes information for titles

- tconst (string) - alphanumeric unique identifier of the title
- averageRating - weighted average of all the individual user ratings
- numVotes - number of votes the title has received

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.

```
SELECT t.titleType, AVG(r.averageRating) AS avg_rating
FROM ratings r
JOIN titles t ON r.tconst = t.tconst
GROUP BY t.titleType
ORDER BY avg_rating DESC;
```

Query Process Results (Status: SUCCEEDED)

Logs

Results

Filter columns...

t.titletype	avg_rating
tvEpisode	7.356931597345818
videoGame	7.1526190476190425
tvMiniSeries	7.141513437057991
tvSeries	6.846716209589187
tvShort	6.81849935316947
tvSpecial	6.759182137481181
tvMovie	6.6394531099556975
video	6.495382761736736
short	6.4263281739093125
movie	6.080975832069286

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 kategórii, ktoré túto podmienku spĺňali a teda predstavujú konkrétne číslo produktov, ktoré boli zohľadnené pri výpočte priemernej hodnoty žánru.

```
SELECT t.titleType, AVG(r.averageRating) AS avg_rating, COUNT(r.numVotes) AS vote_count
FROM ratings r
JOIN titles t ON r.tconst = t.tconst
WHERE r.numVotes > 100
GROUP BY t.titleType
ORDER BY avg_rating DESC;
```

Query Process Results (Status: SUCCEEDED)

[Logs](#)[Results](#)

t.titletype	avg_rating	vote_count
videoGame	7.596835091615776	1801
tvEpisode	7.591775079173472	43891
tvMiniSeries	7.339607602697731	1631
tvSeries	6.950115740740739	8640
tvSpecial	6.858284023668638	1014
tvShort	6.825842696629214	178
short	6.327298019140877	8986
tvMovie	6.1636662868717496	7899
movie	6.039857389124002	63810
video	5.927013972744522	5797

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

```
1 select b.titleType, AVG(r.averageRating) AS avg_rating, SUM(r.numVotes) as vote_count
2 from ratings r join basics b on r.tconst=b.tconst
3 where r.numVotes > 100
4 group by b.titleType
5 order by avg_rating DESC;
```







b.titletype	avg_rating	vote_count
tvEpisode	7.807678244972578	641976
tvSpecial	7.1625	9682
videoGame	7.05	6795
tvMiniSeries	6.9700000000000015	505062
short	6.9275	12828
tvSeries	6.874336283185841	945399
movie	5.9783333333333335	4787783
tvMovie	5.9709677419354845	24832
video	4.566666666666667	433
tvShort	3.1	205

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.

```

1 create view most_popular_movies as
2 select b.tconst,b.primaryTitle,r.averageRating ,r.numVotes
3 from ratings r join basics b on r.tconst=b.tconst
4 where r.numVotes > 100
5 order by r.averageRating DESC,r.numVotes DESC;

```

 default
 foodmart
 moviedb
 basics
 most popular movies
 ratings
[Load more...](#)

most_popular_movies.tconst	most_popular_movies.primarytitle	most_popular_movies.averageRating	most_popular_movies.numvotes
tt10023374	Midnight Sun	9.9	46079
tt10034602	That Day	9.8	39341
tt10116578	Call Me Kevin	9.8	1800
tt10064964	Attack Titan	9.7	23932
tt10008916	The Mongolia Special-Survival of the Fattest	9.7	3791
tt10126480	Watercrazed Corpse Catchers	9.7	116
tt10025702	Halka'nin Avcuklari	9.6	228
tt10064968	The Other Side of the Wall	9.5	20299
tt10084334	A Dark Quiet Death	9.4	4321
tt10042770	Oblivio	9.4	558
tt10022848	Sudu Andagena Kalu Avidin	9.4	419

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 organizáciám umožniť spracovanie obrovského množstva dát rýchlo a efektívne.

Zdroje

<https://datasets.imdbws.com>

[https://www.imdb.com/interfaces/?fbclid=IwAR2UMe3_WL9YpoVLu4GuDY5q7kXai9NIUm9GNp3IFG
KatRrPL7GYP8yk_PM](https://www.imdb.com/interfaces/?fbclid=IwAR2UMe3_WL9YpoVLu4GuDY5q7kXai9NIUm9GNp3IFGKatRrPL7GYP8yk_PM)