

**EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY**

**SPRACOVANIE A ANALÝZA LOGOV V
PRODUKTE HORTONWORKS**

Zadanie z predmetu Big Data

2023

Bc. Ondrej Šima

Bc. Alena Stracenská

Obsah

Úvod	4
1 Hortonworks	5
1.1 Inštalácia Hortonworks Sandbox	5
1.2 Použité dáta	9
1.3 Práca s logmi v Hive	9
1.3.1 Nahranie súboru a udelenie oprávnení	9
1.3.2 Spracovanie a analýza logov	10
Záver	14
Zoznam použitej literatúry	15

Zoznam obrázkov a tabuliek

Obrázok 1	Webová stránka Cloudera, zdroj: [1]	5
Obrázok 2	Obrazovka po inštalácii Hortonworks Sandbox, zdroj: [vlastné spracovanie]	6
Obrázok 3	Obrazovka po zadaní 192.168.109.130:1080, zdroj: [vlastné spracovanie]	6
Obrázok 4	Obrazovka po zadaní 192.168.109.130:4200, zdroj: [vlastné spracovanie]	7
Obrázok 5	Prihlasovacia obrazovka Ambari, zdroj: [vlastné spracovanie] . . .	8
Obrázok 6	Obrazovka Ambari po prihlásení, zdroj: [vlastné spracovanie] . . .	8
Obrázok 7	Ukážka súboru s logmi, zdroj: [vlastné spracovanie]	9
Obrázok 8	Files View v Ambari, zdroj: [vlastné spracovanie]	10
Obrázok 9	Udelenie oprávnení súboru , zdroj: [vlastné spracovanie]	10
Obrázok 10	SQL dopyt pre vytvorenie tabuľky, zdroj: [vlastné spracovanie] . .	11
Obrázok 11	SQL dopyt pre vloženie dát do tabuľky, zdroj: [vlastné spracovanie]	12
Obrázok 12	Vizualizácia rozparsovaných dát v tabuľke, zdroj: [vlastné spracovanie]	12
Obrázok 13	Výsledok SQL dopytu agregácie IP adries, zdroj: [vlastné spracovanie]	13

Úvod

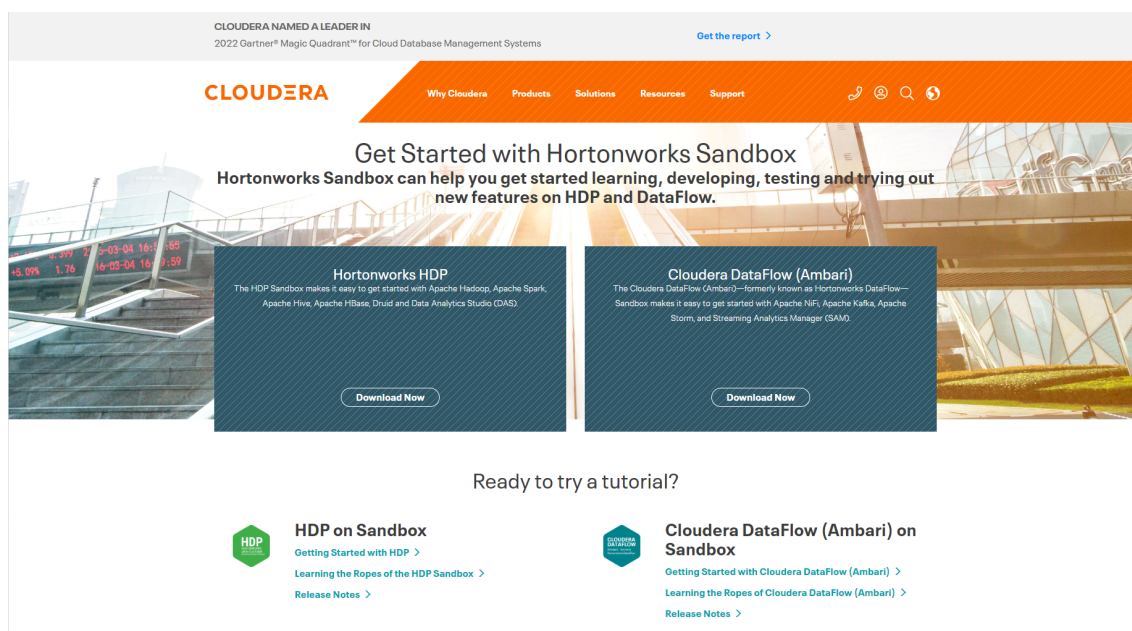
V tomto zadání si ukážeme ako správne nainštalovať Hortonworks na operačný systém Windows. Uvedieme si dôvody prečo sme nepoužili najnovšiu verziu Hortonworks, ale staršiu.

Potom si v krátkosti predstavíme dáta, z ktorých budeme čerpať a ich zdroj a taktiež ich samotné spracovanie a analýzu v už spomenutom produkte.

Veríme, že tento dokument bude prínosný pre čitateľa nie len po teoretickej stránke, ale aj po praktickej, kde sa naučí, čo všetko inštalácia Hortonworks na Windows obnáša a ako je možné spracovať a analyzovať logy v tomto produkte.

1 Hortonworks

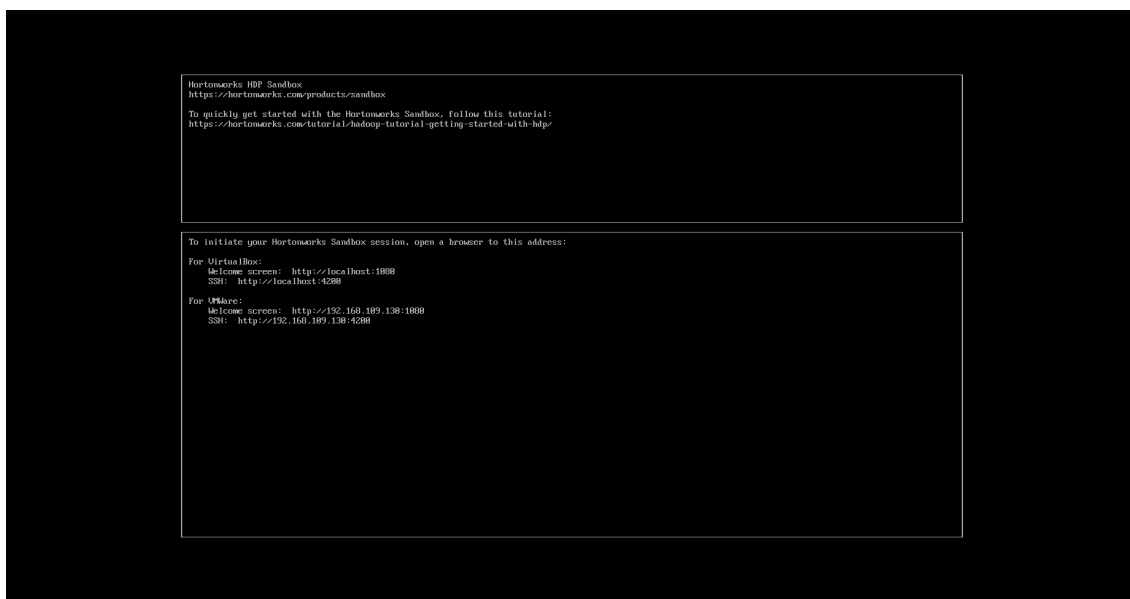
Hortonworks Sandbox predstavuje kontajner pre väčšinu nadstavieb Hadoopu. Je v ňom možné použiť napríklad Hive, Pig a množstvo ďalších doplnkov a nadstavieb. V našom zadaní sme najprv pracovali s verziou HDP_3.0.1 no neskôr sme si kvôli pamäťovej náročnosti a chýbajúcemu Hive View nainštalovali nižšiu HDP_2.6.5 verziu.



Obrázok 1: Webová stránka Cloudera, zdroj: [1]

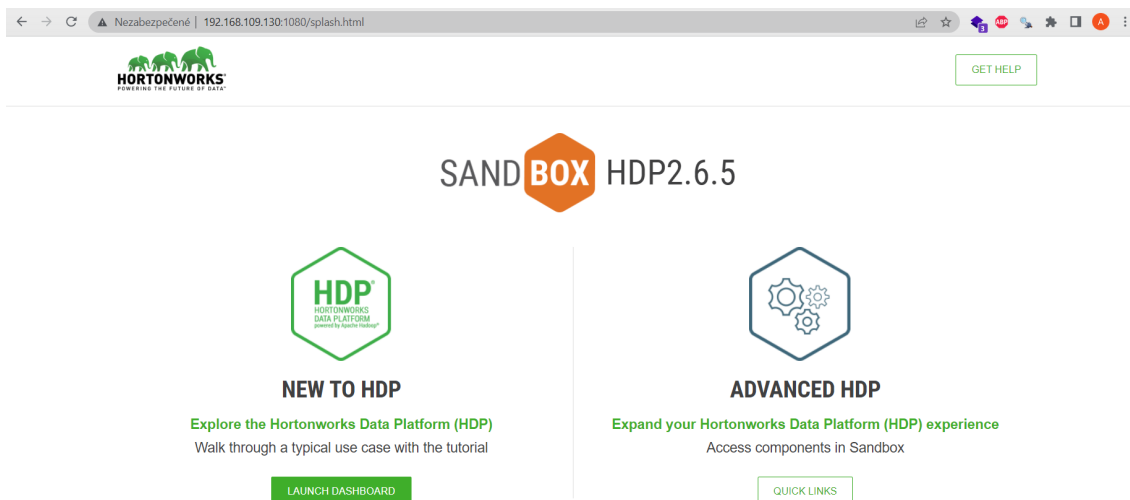
1.1 Inštalácia Hortonworks Sandbox

Najprv si stiahneme z webstránky súbor, v našom prípade sme si stiahli súbor HDP_2.6.5_vmware_180622.ova a po jeho spustení sa nám zobrazila obrazovka s informáciami o inštalácii. Samotná inštalácia trvala približne 15-20 minút. Po inštalácii sa nám zobrazila obrazovka, na ktorej sa nachádzala IP adresa pre zobrazenie Hortonworks Sandbox a webshell samotného Sandboxu. Webshell sme neskôr používali na zmenu root hesla a pridanie nového používateľa pre Ambari.

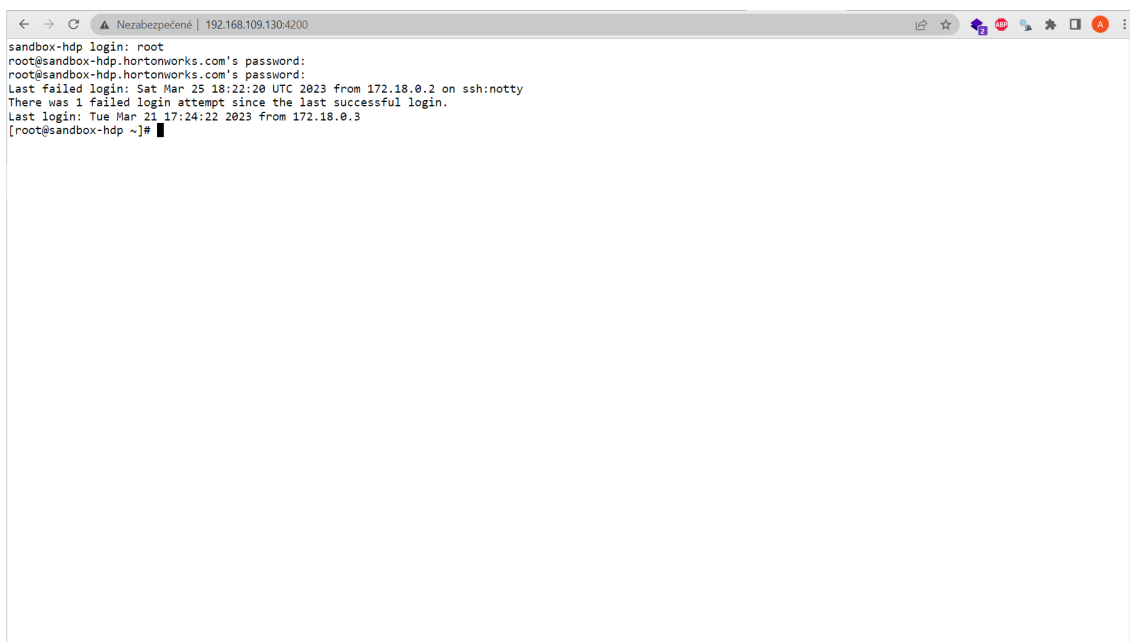


Obrázok 2: Obrazovka po inštalácii Hortonworks Sandbox, zdroj: [vlastné spracovanie]

Po zadaní adries uvedených na obrázku 2, u nás konkrétne na 192.168.109.130:1080 a 192.168.109.130:4200 sa nám zobrazí domovská obrazovka Hortonworks Sandbox a webshell.



Obrázok 3: Obrazovka po zadaní 192.168.109.130:1080, zdroj: [vlastné spracovanie]

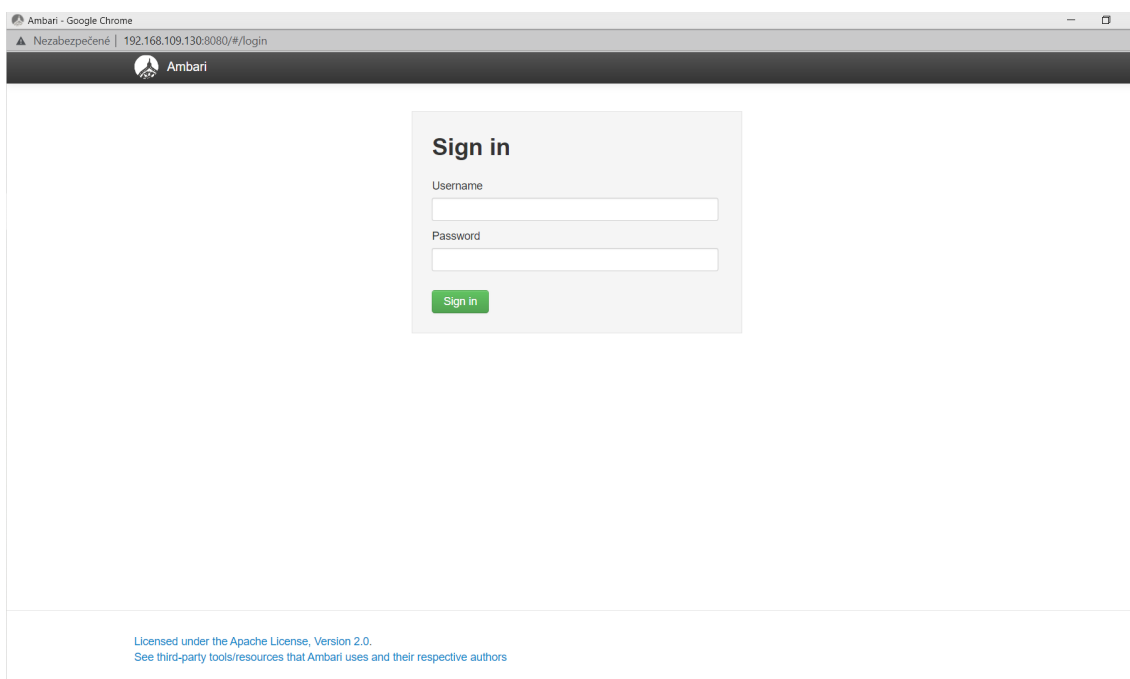


Obrázok 4: Obrazovka po zadaní 192.168.109.130:4200, zdroj: [vlastné spracovanie]

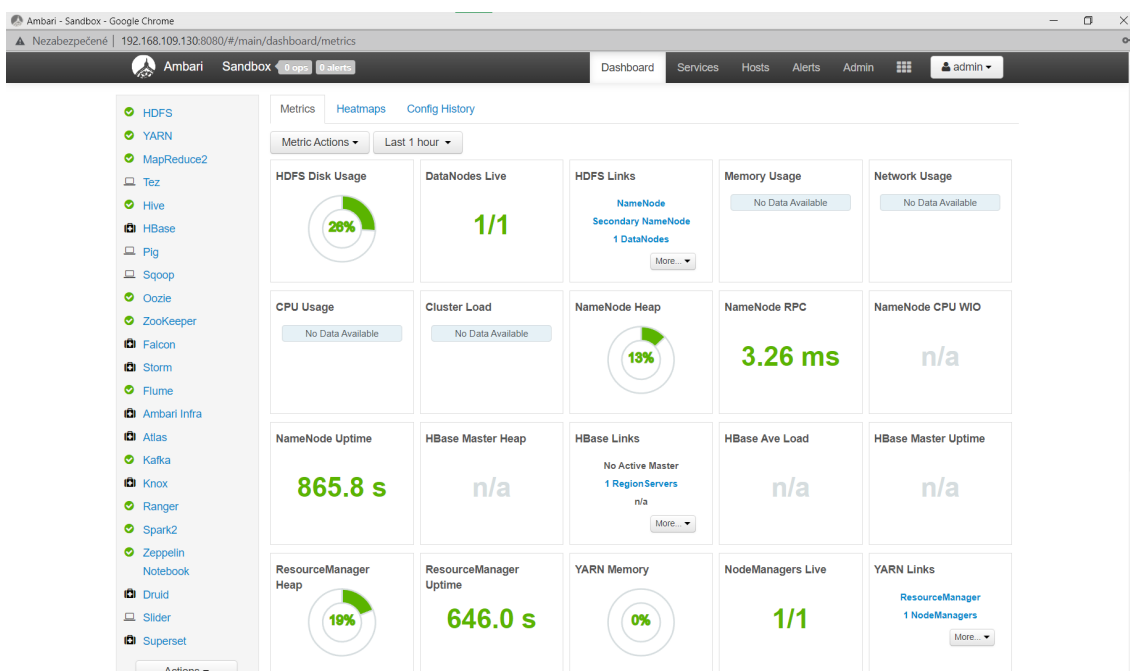
Vo webshelli sa najprv musíme prihlásiť pomocou defaultných prihlasovacích údajov, a to loginu: **root** a passwordu: **hadoop**. Následne dostaneme výzvu na zmenu hesla, ktoré je potrebné zmeniť. Takisto si potom môžeme zmeniť heslo pre admin konto v Ambari a následne sa daným heslom do Ambari prihlásiť. Na prihlasovaciu obrazovku Ambari sa dostaneme po kliknutí na **LAUNCH DASHBOARD** na úvodnej obrazovke Hortonworks Sandbox alebo po zadaní adresy, ktorá je v našom prípade takáto 192.168.109.130:8080.

Následne sa prihlasíme pomocou admin konta a nami nastaveného hesla. Potom sa nám zobrazí obrazovka so všetkými možnými Hadoop nadstavbami a nástrojmi, ktoré môžeme použiť. Služby sa zvyknú spustiť približne do polhodiny. Pohodlnejším spôsobom je pozastavenie celej virtuálky a v prípade potreby jej znovuoobnovenie.

Ak by nám služby nenabehli ani po polhodine, efektívnym spôsobom je ich reštartovanie. To aj v našom prípade pomohlo vyriešiť zopár problémov pri erroroch nadstavby Hive.



Obrázok 5: Prihlasovacia obrazovka Ambari, zdroj: [vlastné spracovanie]



Obrázok 6: Obrazovka Ambari po prihlásení, zdroj: [vlastné spracovanie]

1.2 Použité dáta

Dáta, s ktorými sme pracovali boli vo forme apačovských webových logov. Môžeme si ich stiahnuť na [kaggle.com](https://www.kaggle.com). Sú to pološtruktúrované dáta znázornené na obrázku 7.

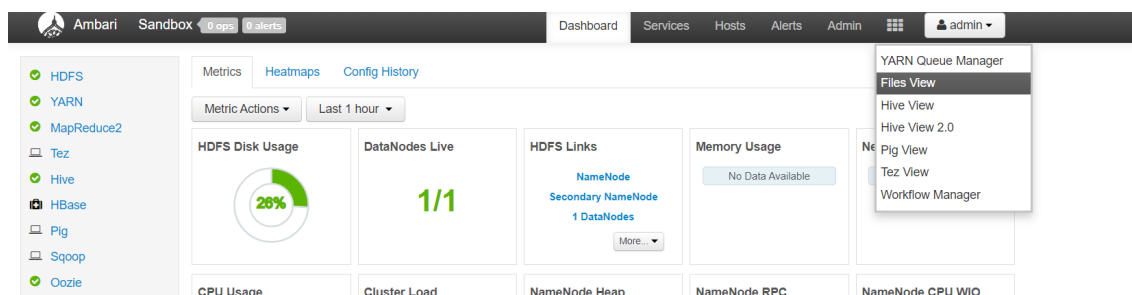
[illegible]

Obrázok 7: Ukážka súboru s logmi, zdroj: [vlastné spracovanie]

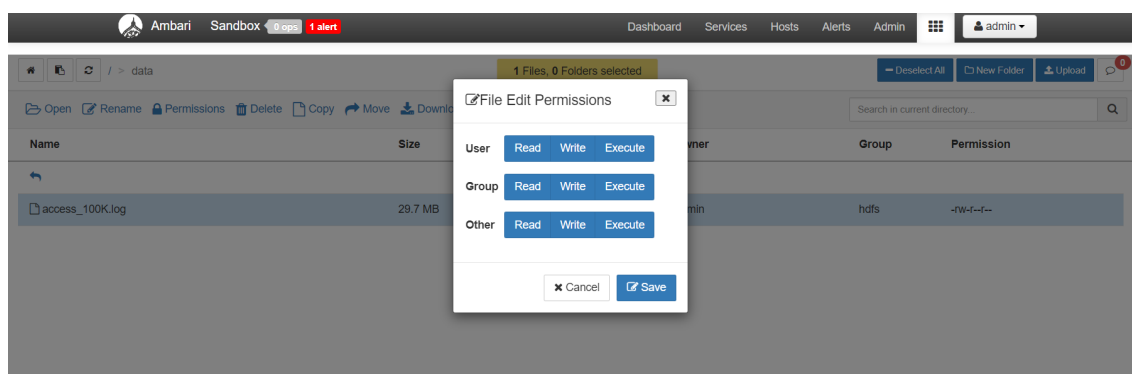
1.3 Práce s logmi v Hive

1.3.1 Nahratie súboru a udelenie oprávnení

Základom bolo nahráť upravený súbor do nami vytvoreného priečinka `data` v Ambari a to konkrétne v záložke `Files View`, súbor sme nazvali `access_100K.log` podľa toho, že sme nebrali viac ako 10 miliónov záznamov, ale len 100 000, kvôli rýchlosti a funkčnosti samotného Ambari. Potom sme mu prideliť všetky oprávnenia `read`, `write` a `execute`. Takisto sme rovnaké oprávnenia udelili aj nami vytvorenému priečinku, v ktorom sa súbor nachádza. Na obrázkoch 8 a 9 môžeme vidieť kde v Ambari nájdeme `Files View` a ako udeľujeme oprávnenia.



Obrázok 8: Files View v Ambari, zdroj: [vlastné spracovanie]



Obrázok 9: Udelenie oprávnení súboru , zdroj: [vlastné spracovanie]

1.3.2 Spracovanie a analýza logov

Do prostredia Hive sme sa prepli rovnakou cestou ako k Files View, ale zvolili sme Hive View. Základom pre spracovanie a analýzu týchto logov bolo vytvoriť si tabuľku, do ktorej sme pomocou regexu parsovali jednotlivé logy. Tabuľku sme rozdelili na 10 polí a pomenovali podľa jednotlivých zložiek. Ak sme tabuľku vopred už mali vytvorenú, tak sme si ju vymazali.

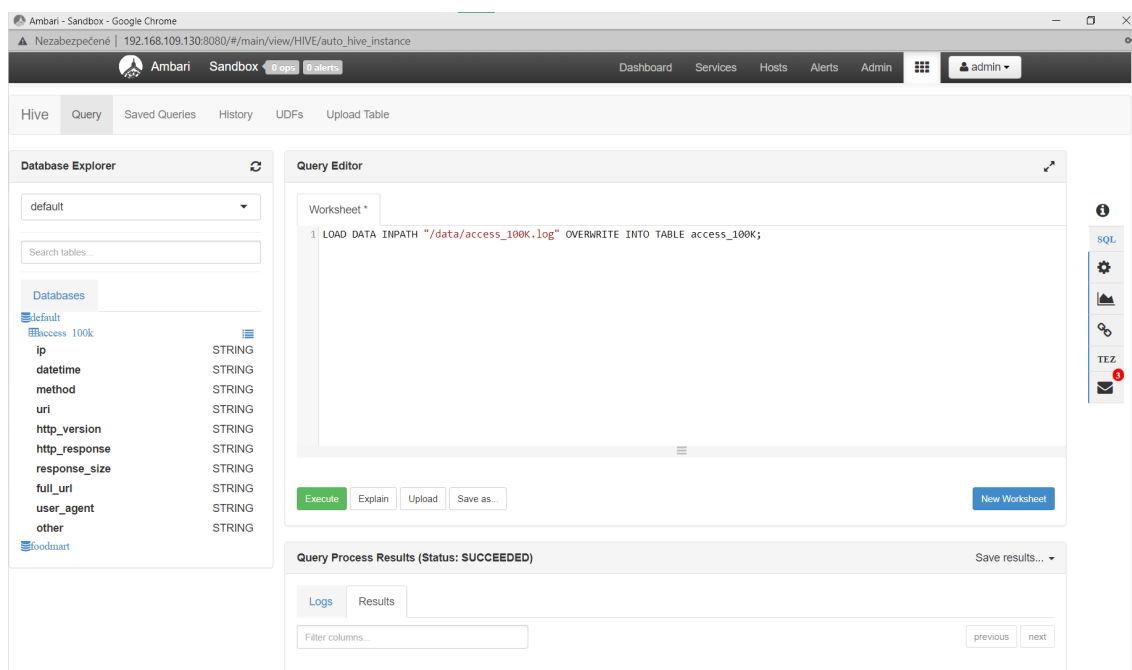
```
DROP TABLE IF EXISTS access_100K PURGE;
```

```
CREATE TABLE access_100K (
  ip STRING,
  datetime STRING,
  method STRING,
  uri STRING,
  http_version STRING,
  http_response STRING,
  response_size STRING,
```

Regex pre členenie logu sme si spravili na <https://regex101.com/r/z9JTRm/1>. Po otvorení si vieme pozrieť pekne jednotlivé vyparované položky logu a pochopiť syntax regexu. Následne sme vyššie uvedený kód vložili do Hive konzoly.



```
LOAD DATA INPATH "/data/access_100K.log" OVERWRITE INTO TABLE access_100K;
```



Obrázok 11: SQL dopyt pre vloženie dát do tabuľky, zdroj: [vlastné spracovanie]

Následne sme si pozreli rozparsované dáta, môžeme ich vidieť na obrázku 12. Sú na ňom znázornené len prvé 4 stĺpce, pre zobrazenie ostatných si musíme posunúť slider na konci obrazovky.

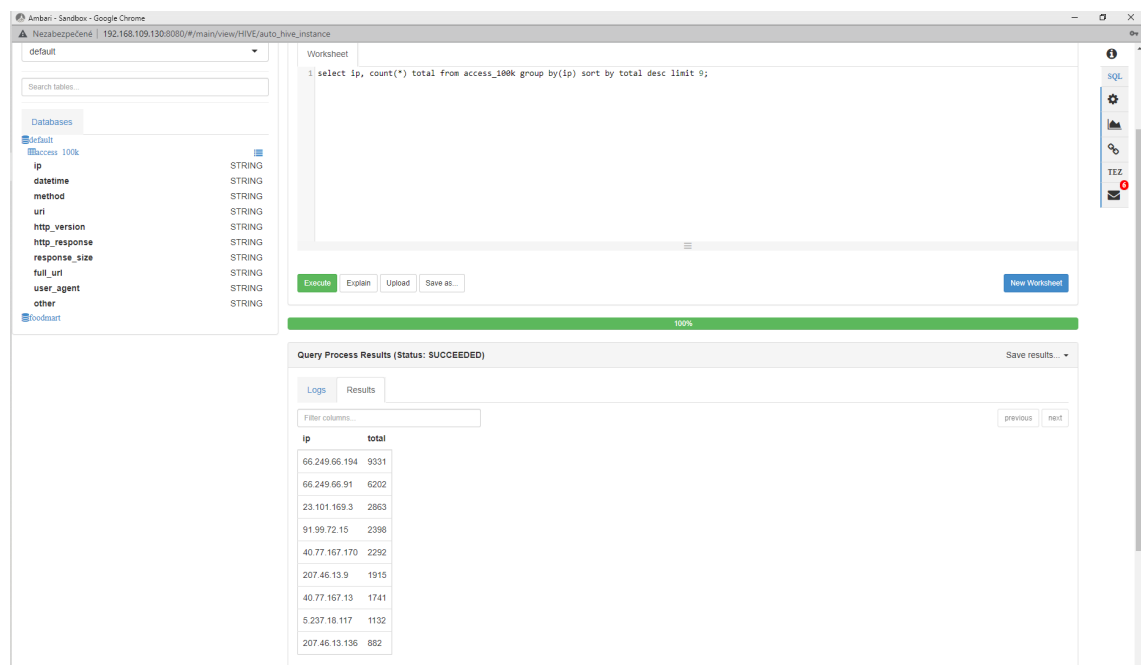
access_100k.ip	access_100k.datetime	access_100k.method	access_100k.uri
54.36.149.41	22/Jan/2019:03:56:14 +0330	GET	/filter/2713%20%D9%85%DA%AF%D8%A7%D9%BE%DB%8C%DA%A9%D8%B3%D
31.56.96.51	22/Jan/2019:03:56:16 +0330	GET	/image/60844/productModel/200x200
31.56.96.51	22/Jan/2019:03:56:16 +0330	GET	/image/61474/productModel/200x200
40.77.167.129	22/Jan/2019:03:56:17 +0330	GET	/image/14925/productModel/100x100
91.99.72.15	22/Jan/2019:03:56:17 +0330	GET	/product/31893/62100%D8%B3%D8%B4%D9%88%D8%A7%D8%B1-%D8%AE%D8%
40.77.167.129	22/Jan/2019:03:56:17 +0330	GET	/image/23488/productModel/150x150
40.77.167.129	22/Jan/2019:03:56:18 +0330	GET	/image/45437/productModel/150x150
40.77.167.129	22/Jan/2019:03:56:18 +0330	GET	/image/576/article/100x100
66.249.66.194	22/Jan/2019:03:56:18	GET	/filter/b41,b665,c150%7C%D8%A8%D8%AE%D8%A7%D8%B1%D9%BE%D8%B2,p5f

Obrázok 12: Vizualizácia rozparsovaných dát v tabuľke, zdroj: [vlastné spracovanie]

Posledným krokom bolo nad tabuľkou spraviť agregácie. Ako príklad si uveďme koľkokrát sa vyskytujú dané IP adresy v súbore resp. môžeme vidieť najaktívnejšie IP adresy. Môžeme tak spraviť prostredníctvom príkazu uvedeného nižšie.

```
select ip, count(*) total from access_100K group by(ip) sort by total desc limit 9;
```

Následne počkáme zopár minút pre zobrazenie výsledku nášho dopytu. V tejto fáze sú volané potrebné súčasti Hadoopu, a to MapReducer, ktorý vykoná svoju úlohu a následne získame výsledok.



The screenshot shows the Ambari web interface in a Google Chrome browser. On the left, a sidebar lists databases and tables, including 'access_100k' with columns like 'ip', 'datetime', 'method', 'uri', 'http_version', 'http_response', 'response_size', 'full_uri', 'user_agent', and 'other'. The main area displays a worksheet with the SQL query: `select ip, count(*) total from access_100k group by(ip) sort by total desc limit 9;`. Below the query editor, a green progress bar indicates 100% completion. The 'Query Process Results (Status: SUCCEEDED)' section shows a table with the following data:

ip	total
66.249.66.194	9331
66.249.66.91	6202
23.101.169.3	2863
91.99.72.15	2398
40.77.167.170	2292
207.46.13.9	1915
40.77.167.13	1741
5.237.18.117	1132
207.46.13.136	882

Obrázok 13: Výsledok SQL dopytu agregácie IP adries, zdroj: [vlastné spracovanie]

Podobným spôsobom by sme pomocou ďalšieho SQL dopytu mohli napríklad rozdeliť čas na jednotlivé zložky, ak by sme chceli zobraziť grafy prípadne iné metriky.

Záver

Cieľom tohto zadania bolo opísať inštaláciu Hortonworks Sandbox a vytvoriť tak návod. Počas inštalácie sme sa stretli s problémami, ktoré spočívali v nedostatku pamäte na našom notebooku a z verzie 3.0.1. sme museli prejsť na verziu 2.6.5. Nižšia verzia, ktorú sme následne aj používali bola plynulejšia a menej nám sekala.

Spracovanie logov sme mohli síce spraviť cez Python, parsovanie pomocou regexov nám prišlo jednoduchšie a rýchlejšie. Samotné vykonávanie dopytov cez SQL nám bolo už vopred známe či už z predmetov na prvom stupni štúdia a taktiež z praxe popri škole.

Veríme, že toto zadanie prinesie čitateľovi nový pohľad na možnosť spracovávať veľké súbory a získa nové prípadne obohatí už doterajšie vedomosti a schopnosti o danej problematike Big Data.

Zoznam použitej literatúry

- [1] CLOUDERA. 2023. *Hortonworks Data Platform (HDP®) on Hortonworks Sandbox*. In cloudera.com [online]. 2023. [citované dňa 25.03.2023]. Dostupné na internete: <https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html>.