**EKONOMICKÁ UNIVERZITA V BRATISLAVE**

**Fakulta hospodárskej informatiky**

Logo

Description automatically generated

**Projekt z predmetu Big Data 01**

Bc. Richard Mišek

2023   Bc. Pavol Melko

Obsah

[Úvod 3](#_Toc131858213)

[1. Počítanie slov lokákne na počítači 4](#_Toc131858214)

[2. Počítanie slov použitím Hadoop 7](#_Toc131858215)

[Záver 10](#_Toc131858216)

[Zoznam príloh 11](#_Toc131858217)

# Úvod

Úlohou tohto zadania je vymyslieť textový súbor s nejakým obsahom a následne pomocou Cloudera VM použiť Hadoop na zistenie výskytu každého slova v texte. Inými slovami sa pokúsime spočítať koľkokrát sa každé slovo nachádza v texte. Budeme si musiet nainštalovať Cloudera VM, ktorý už má v sebe zabudovanú kompletnú funkcionalitu Hadoopu. Konkrétne sa jedná o jeho distrubúciu. Následne si vytvoríme textový dokument a potom postupne prejdeme prípravov až následne spustíme samotný program a uvidíme či nám vypíše správne výsledky. Správnosť výsledkov budeme porovnávať voči lokálne spustenej aplikácií.

# Počítanie slov lokákne na počítači

V prvej časti tohto zadania si vytvoríme textový dokument s náhodnými vetami. Pomocou textového editora vim a nasledovných príkazov sme vytvorili 3 jednotlivé súbory. Nižšie na obrázku 1 môžeme vidieť príkaz, pomocou ktorého sme vytvorili textový súbor a na obrázku 2 jeho konkrétny obsah. Jedná sa kompletne iba o náhodný text.



Obrázok 1: Vytvorenie textového súboru

Text, letter

Description automatically generated

Obrázok 2: Obsah textového súboru



Obrázok 3: Vytvorenie súboru mapper.py

Nižšie na obrázku 3 môžeme vidieť kód, ktorý sme vložili do súrobu mapper.py. Jedná sa o programovací jazyk python, o čom svedčí aj prípona *py.* V tomto kóde si importujeme knižnicu sys na používanie STDIN a STDOUT. Ďalej kód odstaňuje biele znaky, rozdeľuje riadky na slová a v poslednom rade nasleduje samotný výpis jednotlivých slov.

**Poznámka:** pri používaní kódu sme museli zmeniť apostrofy na úvodzovky.

Text

Description automatically generated

Obrázok 4: Kód vložený do súboru mapper.py

Ďalším krokom bolo vytvoriť súbor reducer.py a vložiť doňho potrebný kód. Nižšie na obrázkoch 5 a 6 môžeme vidieť príkaz na vytvorenie súboru a aj samotný kód. Tento kód nám už počíta jednotlivé slová. Využíva sa tu cyklus for kde pre každý riadok a každé slovo priráta 1 počet.

**Poznámka:** Takisto sme museli v použitom kóde prepísať klasické apostrofy na dvojité úvodzovky. Inak to hádzalo syntaktickú chybu.



Obrázok 5: Vytvorenie súboru reducer.py

Graphical user interface, text, application

Description automatically generated

Obrázok 6: Kód v súbore reducer.py

Posledným krokom tejto časti bolo spustiť samotný program. Príkaz na spustenie môžeme vidieť na obrázku 7. Zároveň môžeme vidieť aj výstup zo súboru a počet jednotlivých slov, ktoré sa nachádzajú už v skôr spomenutom textovom súbore.

Graphical user interface, text

Description automatically generated with medium confidence

Obrázok 7: Lokálny výpis

# Počítanie slov použitím Hadoop

Prvým krokom bola inštalácia Cloudery vo virtualboxe. Nainštalovali sme teda novú virtuálku, ktorá už obsahovala v sebe cloudera inštaláciu. Ako operačný systém sme použili Linux s distribúciou RedHat alebo inak aj Centos. Inštalácia prebehla úspešne a výsledok môžeme vidieť na obrázku 8.

Graphical user interface, application

Description automatically generated

Obrázok : Prihlasovací formulár do Cloudera

**Pri prvom prihlásení sme mali problémy, pretože sme dostávali error: Unable to connect. Tento problém bolo veľmi jednoduché vyriešiť len reštartovaním celého systému. Príkaz na reštartovanie nájdeme nižšie na obrázku 9.**



Obrázok : Príkaz na reštartovanie systému

V ďalšom kroku sme po prihlásení sa do systému spustili YARN a HDFS servis. Všetko prebehlo bez komplikácií a nepotrebovali sme nič debugovať. Ďalej sme použili kód na vytvorenie adresára v hdfs. Kód môžeme vidieť na obrázku 10.



Obrázok : Vytvorenie adresara pre nas textovy subor v hdfs

Cieľom tohto je aby sme mali miesto, kde budeme vedieť neskôr prekopírovať náš textový súbor. Nižšie na obrázku 11 sme si overili či bol adresár správne vytvorený. Náš adresár je podčiarknutý červenou farbou.

Table

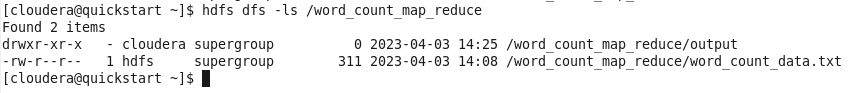
Description automatically generated

Obrázok : overenie či sa adresár vytvoril

Na obrázku 12 môžeme vidieť kód, pomocou ktorého sme nakopírovali náš textový súbor do nášho novovytvoreného adresára. Pod ním na obrázku 13 môžeme vidieť overenie, že náš textový súbor sa naozaj nachádza v nami vytvorenom adresári na hdfs.



Obrázok : kopírovanie textového súboru do nášho nového adresára



Obrázok : Overenie či sa súbor skopíroval

Ďalším krokom je zmena oprávnení pre súbory mapper.py a reducer.py tak, aby clodera účet mal prístup k nim a zároveň aj k nášmu adresáru na hdfs. Použité príkazy môžeme vidieť na obrázku 14.



Obrázok : Pridelenie oprávnení na jednotlivé python súbory

Blížime sa k posledným dvom krokom tohto zadania. Predposledný krok je stiahnutie hadoop-streaming jar súboru z oficiálnej stránky. Následne treba tento súbor extrahovať do nášho domovského adresára. Raz, keď už je toto hotové, tak spustíme nasledovný príkaz, ktorý je viditeľný na obrázku 15. Je dôležité aby v tomto príkaze boli posledné dva riadky začínajúce príkazom file. Tento príkaz pridáva referenciu na súbory reducer.py a mapper.py a teda tieto súbory budú použité pri samotnom procese. Výsledkom bude output v nami vytvorenom adresári v hdfs. Na obrázku 16 môžeme vidieť posledný použitý príkaz, pomocou ktorého si vypíšeme výsledky z outputu.

Text

Description automatically generated

Obrázok : Spustenie JAR súboru

Graphical user interface, application

Description automatically generated with medium confidence

Obrázok : Výsledok od Hadoop

# Záver

Cieľom tejto práce bolo naučiť sa robiť s cloudera nástrojom, ktorý má už v sebe zabudovaný hadoop, a zároveň pripraviť textový súbor a kód, ktorý nám bude tento textový súbor čítať a počítať jednotlivé slová v ňom. Výsledkom je fungujúci cloudera nástroj a takisto aj fungujúci kód, ktorý nám spočíta jednotlivé slová v textovom súbore a vypíše ich do konzoly.

# Zoznam príloh

[Obrázok 1: Vytvorenie textového súboru 4](#_Toc131858169)

[Obrázok 2: Obsah textového súboru 4](#_Toc131858170)

[Obrázok 3: Vytvorenie súboru mapper.py 4](#_Toc131858171)

[Obrázok 4: Kód vložený do súboru mapper.py 5](#_Toc131858172)

[Obrázok 5: Vytvorenie súboru reducer.py 5](#_Toc131858173)

[Obrázok 6: Kód v súbore reducer.py 5](#_Toc131858174)

[Obrázok 7: Lokálny výpis 6](#_Toc131858175)

[Obrázok 8: Prihlasovací formulár do Cloudera 7](#_Toc131858176)

[Obrázok 9: Príkaz na reštartovanie systému 7](#_Toc131858177)

[Obrázok 10: Vytvorenie adresara pre nas textovy subor v hdfs 7](#_Toc131858178)

[Obrázok 11: overenie či sa adresár vytvoril 8](#_Toc131858179)

[Obrázok 12: kopírovanie textového súboru do nášho nového adresára 8](#_Toc131858180)

[Obrázok 13: Overenie či sa súbor skopíroval 8](#_Toc131858181)

[Obrázok 14: Pridelenie oprávnení na jednotlivé python súbory 8](#_Toc131858182)

[Obrázok 15: Spustenie JAR súboru 9](#_Toc131858183)

[Obrázok 16: Výsledok od Hadoop 9](#_Toc131858184)