Hadoop for Big data

Készítette: Madarász Zsolt

Szak: PTI

Dátum: 2019.05.31.

Tartalom

[2 Bevezetés 2](#_Toc10239775)

[3 Mi az a Hadoop? 2](#_Toc10239776)

[3.1 A Hadoop fő részei: 3](#_Toc10239777)

[3.2 A Hive 3](#_Toc10239778)

[4 Telepítés 3](#_Toc10239779)

[4.1 A szükséges szoftverek letöltése 3](#_Toc10239780)

[4.2 Környezet telepítése (https://hortonworks.com) 3](#_Toc10239781)

[5 A project 4](#_Toc10239782)

[5.1 Leírás 4](#_Toc10239783)

[5.2 A megvalósítás lépései 4](#_Toc10239784)

[6 Források 8](#_Toc10239785)

[**1. ábra** A Hive elérése SSH-n keresztül 4](#_Toc10240099)

[**2. ábra** . csatlakozás PUTTY-al a Hive-hoz 4](#_Toc10240100)

[**3. ábra** Mozilista letöltése 5](#_Toc10240101)

[**4. ábra** Mozilista kitömörítése 5](#_Toc10240102)

[**5. ábra** az item és user elemek felmásolása 6](#_Toc10240103)

[**6. ábra** az item és user elemek felmásolása 6](#_Toc10240104)

[**7. ábra** A tábla kilistázása 6](#_Toc10240105)

[**8. ábra** Adatok feltöltése a Movies táblába 7](#_Toc10240106)

[**9. ábra** Szűrés Map reduce használatával 7](#_Toc10240107)

# Bevezetés

Azért választottam ezt a témát, mert érdekelnek az adatbázisos rendszerek és a nagy adathalmazok kezelése, mint megoldandó probléma. A Hadoop-ról még soha nem hallottam előtte, így kíváncsian indultam neki a feladatnak. Sose rossz, ha tanulhatok valami ujjat. A feladat Hadoop segítségével feldolgozni nagy mennyiségű adathalmazt, majd kinyerni belőle számunkra fontos információkat. Az információk kinyeréséhez Hive-ot használtam.

# Mi az a Hadoop?

Az **Apache Hadoop** (Wikipedia, 2017) egy nyílt forráskódú keretrendszer, amely adat intenzív elosztott alkalmazásokat támogat. Nagy mennyiségű alacsony költségű, általánosan elérhető hardverből épített szerverfürtök építését teszi lehetővé. A Hadoop a Google MapReduce és a Google File System leírásaiból készült. A Hadoop projektet Doug Cutting és Mike Cafarella hozta létre 2005-ben. Cutting, aki akkor a Yahoo!-nál dolgozott, fiának játék elefántjáról nevezte el a projektet.

## A Hadoop fő részei:

* **Hadoop Common**
* amely a fájlrendszerrel és a operációs rendszerrel kapcsolatos absztrakciókat tartalmazza, valamint azokat a szkripteket és programokat, amelyek a Hadoop rendszer indításához szükségesek
* **MapReduce rendszer**
* A **MapReduce** egy programozási modell nagy adathalmazok feldolgozására párhuzamosan és egy szerverfürtön elosztottan.
* A MapReduce tartalmaz egy **map** funkciót, amely szűrést és rendezést végez, valamint egy **reduce** funkciót, amely összegzi az eredményt. A MapReduce rendszer osztja el a feladatokat a szervereken párhuzamosan futtatva azokat, irányítva minden adatátvitelt, egyúttal hibatűrést is biztosít redundancián keresztül.
* A modellt a funkcionális programozásból ismert map és reduce funkciók inspirálták, bár a használatuk nem egészen ugyanaz a MapReduce rendszerben, mint az eredeti formában.
* **Hadoop Distributed File System** (Hadoop elosztott fájlrendszer)
  + A HDFS (Hadoop Distributed File System - Hadoop Elosztott Fájlrendszer) egy elosztott, skálázható és hordozható fájlrendszer, amelyet a Hadoop rendszerhez írtak Java nyelven.

## A Hive

A Hive (IBM, 2014) egy keretrendszer, amit a Hadoop köré terveztek megkönnyítve ezzel a munkánkat. A Hive-ban lehetőségünk nyílik az adatokat adatbázisokban tárolni. Ezeket SQL utasításokkal kezelhetjük. A Hadoop adminisztrációt pedig megteszi helyettünk a Hive. Olyan mintha egy fordító lenne a géphez.

# Telepítés

## A szükséges szoftverek letöltése

A rendszerkörnyezet felállításához szükség van három szoftverre. A virtualboxra, ami emulálja a hadoop környezetet és magára a hadoop-ra a HIVE-al és a többi hadoop-os kiegészítővel. Ezen környezet eléréséhez Putty-ot használunk

**Hadoop Hive-al telepítő link:**

<https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html>

**Virtualbox telepítő link:**

<https://www.virtualbox.org/wiki/Downloads>

**Putty telepítő link:**

<https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html>

## Környezet telepítése

1. Virtualbox feltelepítése (https://hortonworks.com)
2. Hortonworks implementálása: **File -> Import Appliance, itt image file kiválasztása és open.**
3. **Hortonworks elindítása a start-al**
4. **A bootolás után be kell állítani a memóriaméretet, ami 22GB System -> Motherboard -> Base Memory -> OK**
5. **A Hadoop használatához SSH-n keresztül kell csatlakozni, ehhez először telepíteni kell a** Connected Data Architecture (**CDA**)-t. Ehhez a következő sort kell begépelni: **ssh root@sandbox-hdp.hortonworks.com –p 2200**
6. **Script futtatása: cd /sandbox/deploy-scripts/**
7. **majd a script futtatása: sh enable –vm –cda.sh**

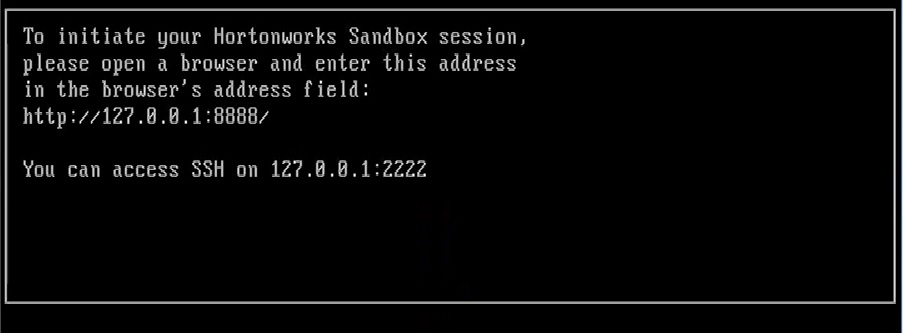
# A project

## Leírás

A feladat Hadoop alkalmazásával nagy adatmennyiség feldolgozása. Egy, vagy több nagyméretű állomány letöltése és a file tartalmának feldolgozása a Hadoop architektúra szerint. Jelen esetben a Hive rendszer használatával. A Hive adatbázisba gyűjti az adatokat és azután egyszerű SQL hívásokkal lehet kinyerni a kívánt információt.

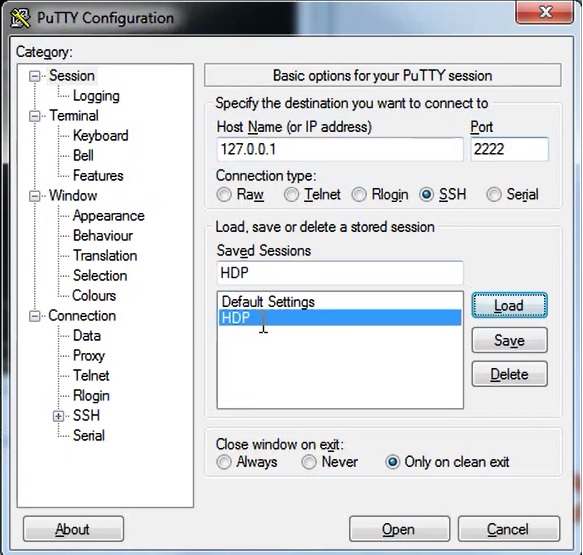
## A megvalósítás lépései

1. Miután a Virtualbox befejezte a Hadoop rendszer telepítését az **1. ábra** mutatja, mit kell látnunk. Ha hasonló képet kapunk, akkor sikeres volt a telepítés. A csatlakozáshoz itt található az IP cím és a port.



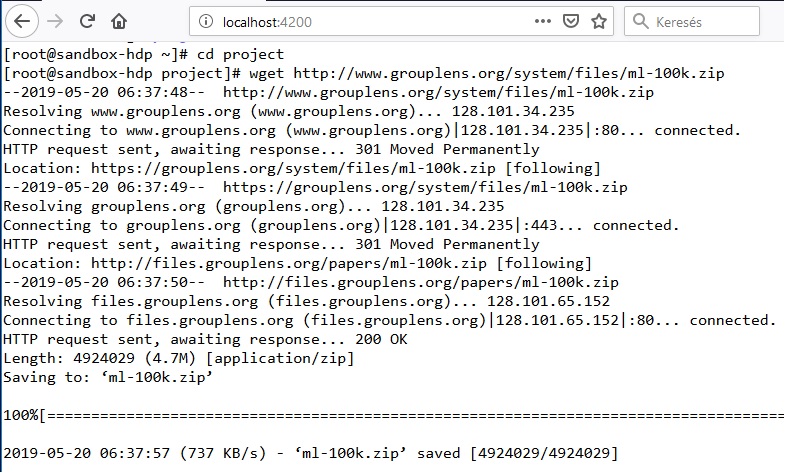
**1. ábra** A Hive elérése SSH-n keresztül

1. Miután kinéztük az IP címet nyissuk meg a Putty programot. Kitölteni értelemszerűen kell. Az IP címhez a címet írjuk a porthoz pedig a portot. Ezt mutatja a **2. ábra**.



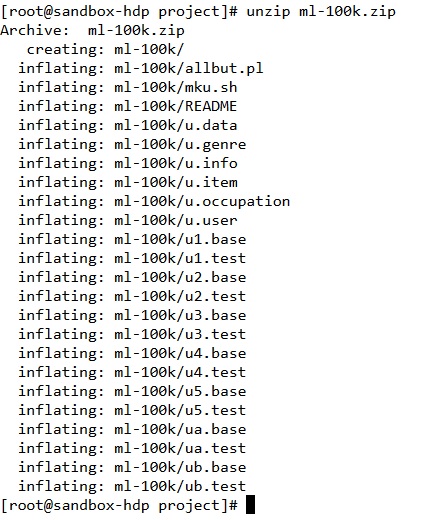
**2. ábra** . csatlakozás PUTTY-al a Hive-hoz

1. A Hadoop Linux alapú, így linuxos parancsokat kell használni. Én létrehoztam egyproject könyvtárat. Ebbe letöltöttem a projekthez használatos fájlokat. Mivel mozilistát elemzünk, így egy hivatalos mozilistát töltöttem le a **WGET** paranccsal. Az eredmény és a pontos parancsot a **3. ábra** szemlélteti.

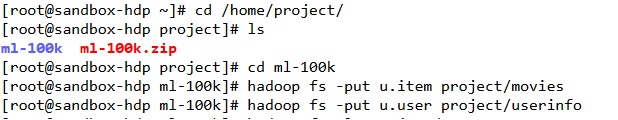


**3. ábra** Mozilista letöltése

1. Letöltés után a file tartalmát ki kell tömöríteni. Ezt szimplán az UNZIP paranccsal megtehetjük. A **4. ábra** mutatja a kitömörített fájlokat. Nekünk csk két file kell. Az item tartalmazza a mozifilmeket kategóriák szerint, az user pedig felhasználok adatait. Ezt a két filet fel kell másolni a Hadoop filerendszerbe. Ez a PUT paranccsal elvégezhető. Ezt szemlélteti az **5. ábra**.



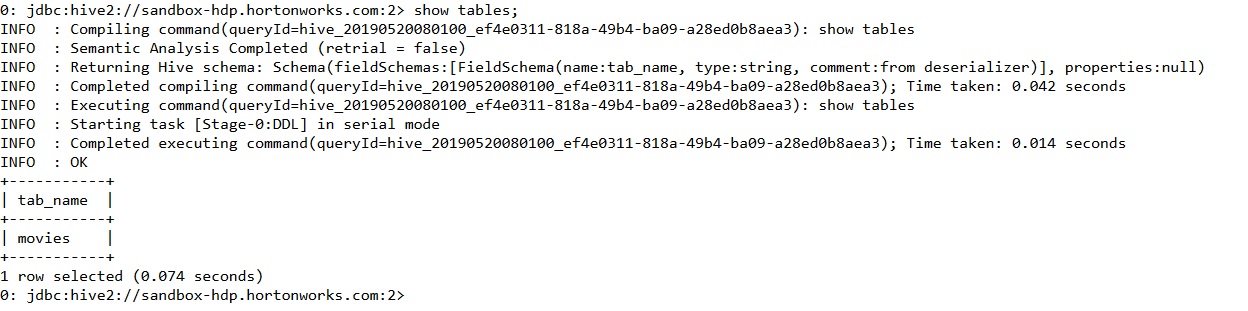
**4. ábra** Mozilista kitömörítése



**5. ábra** az item és user elemek felmásolása

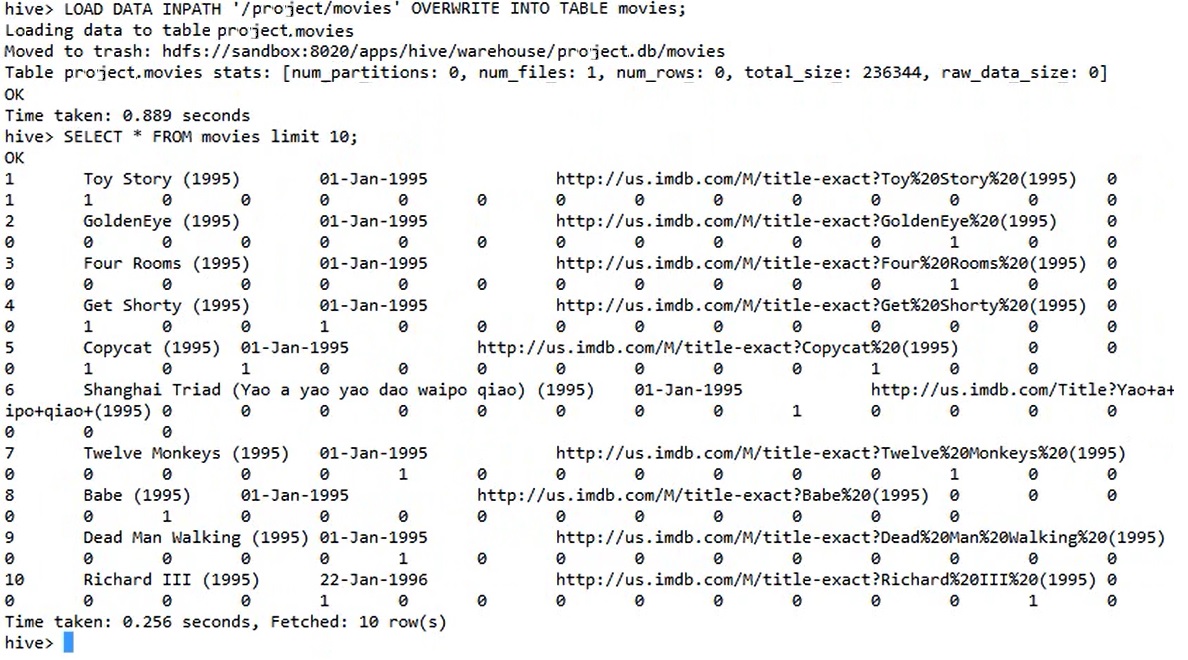
1. Ha kész vagyunk a másolással ideje elindítani a HIVE-ot. Ezt egyszerűen a **hive** begépelésével tehetjük meg. Innentől kezdve SQL utasításokat is kiadhatunk. Ha Hadoop parancsokat szeretnénk futtatni, a parancsok elé ki kell írnunk a !. Innen tudja a fordító, hogy nem Hive-os SQL parancsok fognak futni. Létre kell hoznunk egy adatbázist a CREATE DATABASE adatbázisnév paranccsal. Ezt követően ki kell választani az adatbázist. Ezt az USE paranccsal tehetjük meg. USE adatbázisnév . Most már létrehozhatjuk a táblákat. Az adatbázis kiválasztását és a Movies tábla létrehozását a **6. ábra** szemlélteti. A tábla SQL parancsai megtalálhatóak a **GIT** mellékleten. A létrehozott táblát a SHOW TABLES paranccsal tudjuk lekérdezni. A kilistázást szemlélteti a **7. ábra**

**6. ábra** az item és user elemek felmásolása



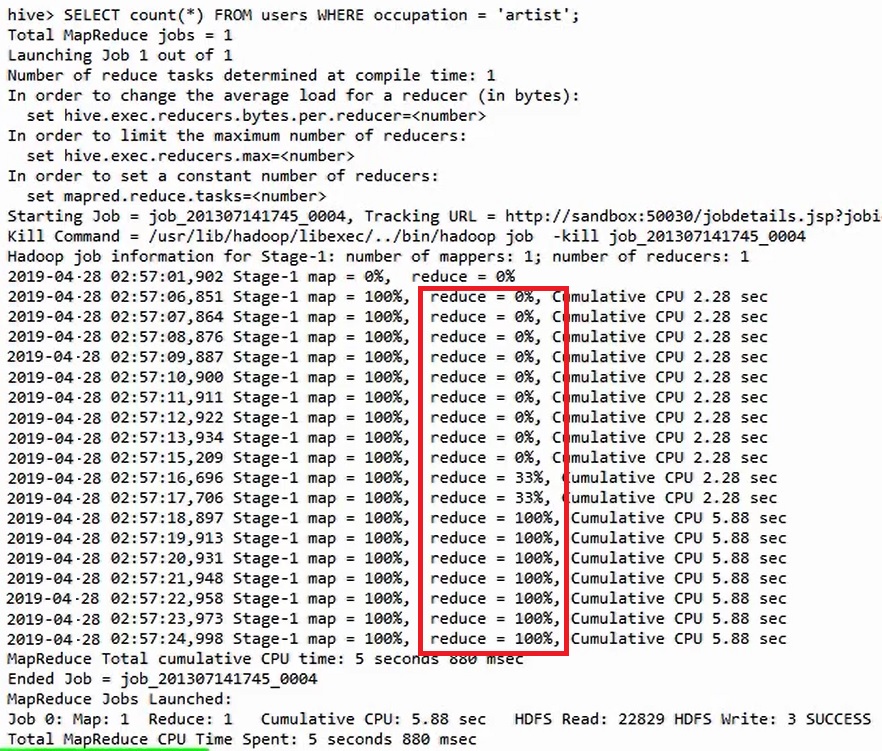
**7. ábra** A tábla kilistázása

1. A táblázat létrehozása után nincs más dolgunk, mint az item file tartalmát belerakni az adatbázis Movies táblájába.Az item a Movies mappába lett belerakva a PUT paranccsal. Ezt a LOAD DATA INPATH ’/project/movies’ INTO TABLE MOVIES paranccsal tehetjük meg. Ha jól csináltuk a táblánk megtelik adattal. Ezt szemlélteti a **8. ábra**



**8. ábra** Adatok feltöltése a Movies táblába

1. Ha ezzel kész vagyunk hozzuk létre a táblát az userinfonak is. A kód megtalálható az SQL fájlban. Töltsük be az adatokat a LOAD DATA paranccsal az userinfóba. Semmi más dolgunk nincs, mint elkezdeni az adatszűrést a SELECT parancsokkal. Egyszerű lekérdezés esetén a Hive nem alkalmazza a Map reduce eljárást, de egy összetett lekérdezés esetén már igen. Ezt szemlélteti a **9. ábra**. Itt jól látható, hogy mivel ez egy összetettebb listázás a Hive utasítja a Hadoop-ot a mapreduce bekapcsolására. Ennek folyamata jól nyomonkísérhető a pirossal bekeretezett területen.



**9. ábra** Szűrés Map reduce használatával

# Források

# Hivatkozások

Github. (2019. 05 31). *Github*. Forrás: Github: https://github.com/zsoltix83/Hadoop.git

*https://hortonworks.com.* (dátum nélk.). Forrás: https://hortonworks.com/tutorial/sandbox-deployment-and-install-guide/section/1/.

IBM. (2014). *Apache.org*. Forrás: Hive apache: https://hive.apache.org/

Pluralsight. (dátum nélk.). *Pluralsight*. Forrás: Pluralsight: www.pluralsight.com

Stackoverflow. (2019). *Stackoverflow*. Forrás: Stackoverflow: https://stackoverflow.com/

Wikipedia. (2017. augusztus 27). *wikipedia*. Forrás: https://hu.wikipedia.org/wiki/Apache\_Hadoop