# Assignment 1

Prerequisities:

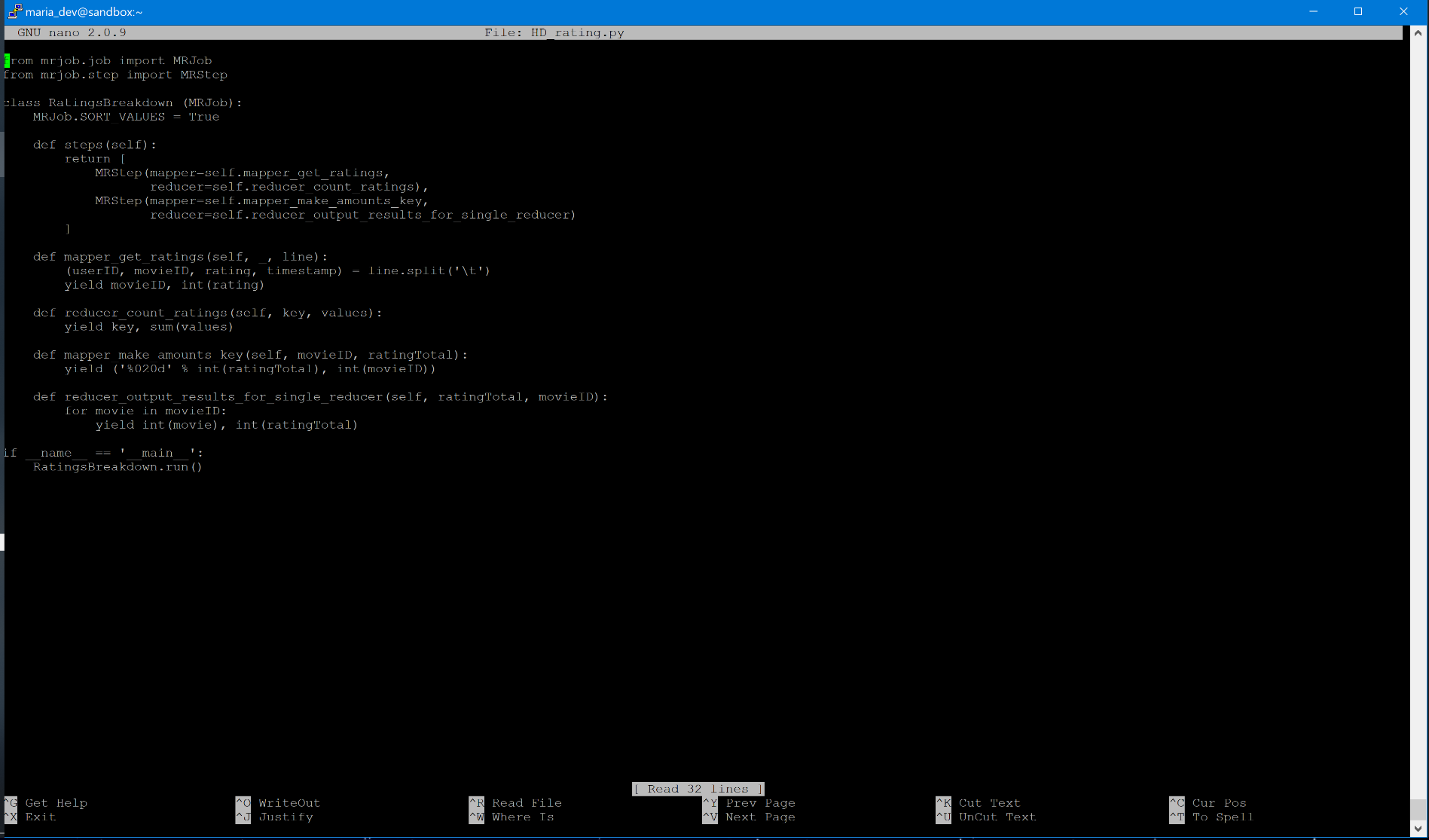
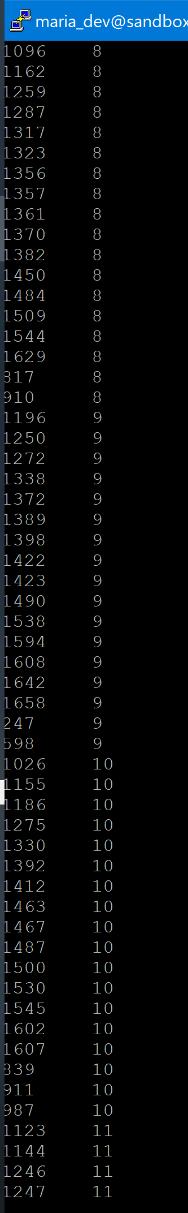
1. Installeer VirtualBox
2. Installeer HortonWorks sandbox versie 2.5.0
3. Importeer sandbox in VirtualBox
4. Download MovieLens dataset
5. Start sandbox in VirtualBox

Ondernomen stappen:

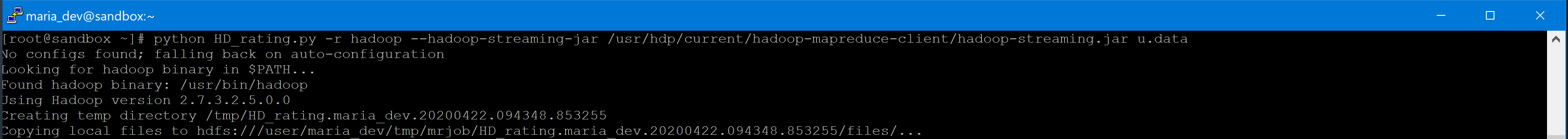
1. Start PuTTY en login op 127.0.0.1:2222 (maak een commandline sessie aan)
2. su root (verkrijg admin/root-rechten)
3. cd /etc/yum.repos.d (ga naar map /etc/yum.repos.d)
4. cp sandbox.repo /tmp (kopieer sandbox.repo naar de folder tmp)
5. rm sandbox.repo → y (verwijder sandbox.repo)
6. cd ~ (ga naar de rootfolder)
7. yum install python-pip (installeer client om packages te downloaden)
8. pip install google-api-python-client==1.6.4 (installeer benodigde packages)
9. pip install mrjob==0.5.11 (installeer benodigde packages)
10. yum install nano (installeer commandline code editor)
11. wget <http://witan.nl/hadoop/u.data> (download u.data bestand)
12. wget <http://witan.nl/hadoop/u.item> (download u.item bestand)
13. wget <http://witan.nl/hadoop/HD_rating.py> (download HD\_rating.py bestand)
14. nano HD\_rating.py (edit HD\_rating.py bestand)
15. *Python-script schrijven (zie* [*https://www.github.com/LennartdeWaart/Hadoop*](https://www.github.com/LennartdeWaart/Hadoop)*)*
16. CTL + X → y → enter (sluit bestand en sla op)
17. python HD\_rating.py -r hadoop --hadoop-streaming-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar u.data (run HD\_rating.py bestand met behulp van Hdp FileSystem)

*Uitleg code in script. Zie onderstaande URL:* <https://github.com/LennartdeWaart/Hadoop/blob/master/Assignment_1/break_down_ratings.py>

**Resultaat**

Omdat er een oudere versie (2.6) van Python wordt gebruikt in Hadoop lukt het niet om het volledige script werkende te krijgen. Lokaal, waar ik de laatste versie van Python geïnstalleerd heb, werkt deze echter wel. Ik heb getracht de laatste versie van Python op Hadoop te installeren, maar dit lukt mij helaas niet. Om te bewijzen dat ik het script (weliswaar vereenvoudigd) wel kan laten draaien in Hadoop voeg ik screenshots toe van het script dat wel aan de 6-beoordeling voldoet maar niet aan de 8-beoordeling. Daaronder voeg ik een screenshot van het lokale script en haar resultaten toe.

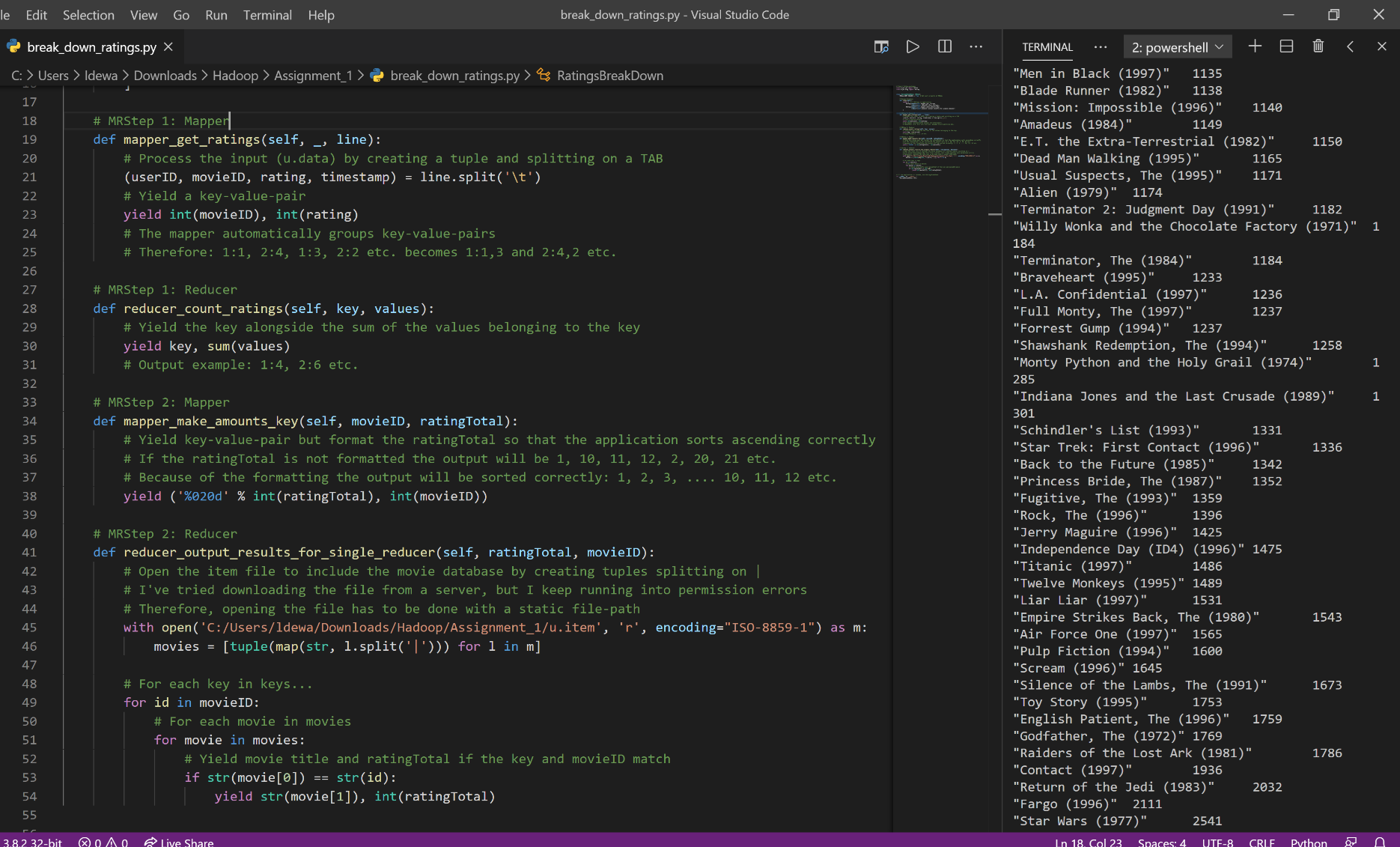
**Schermopname 1: Python script (vereenvoudigd) in nano PuTTY**



**Schermopname 2: Python script (vereenvoudigd) aangeroepen via Hadoop**

**Schermopname 3: Python script (vereenvoudigd) uitgevoerd**

Hier staan dus de movieIDs gesorteerd op hun totale rating

**Schermopname 4: Python script (8-rating) lokaal uitgevoerd**

Wanneer dit script lokaal wordt uitgevoerd (met de laatste versie van Python) werkt deze naar de 8-rating behorende.

De namen van de movies worden getoond in combinatie met hun totale rating (gesorteerd).