Spark og HDFS

Gruppe 13: Mats Engelien og Lars Erik Faber

# Lese CSV filer

Før vi kan begynne å lage aggregeringer, må vi gjøre datasettene om til Parquet filer. Det er alltid lurt å gjøre om til Parquet først, siden de tilbyr komprimering.

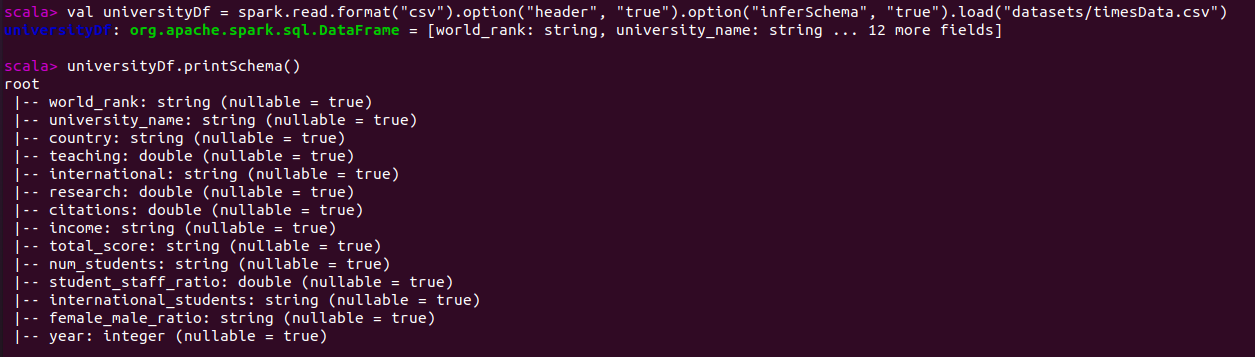
## Student Performance Dataset (KV-database)

asd

## Socio-Economic Country Profiles (Dokumentdatabase)

asd

## World University Rankings (Kolonnefamiliedatabase)

Først leser vi csv filen med «inferSchema» satt til «true» for å opprette schema.

Jeg legger merke til at noen av datatypene ikke stemmer. Jeg ser at «world\_rank» er satt til streng, selv om det tilsynelatende er int. Men det går fin, siden den også inneholder verdier som «100-150» lenger ned i datasettet. Derimot, «international» var satt til streng, fordi null-verdier var byttet ut med

## Government Types Of The World (Grafdatabase)

asd

# Endringer på nettsiden

På grunn av endringene i komponentene …

# HDFS

asd

