**Storing on Hadoop**

1. I have installed Hadoop on an Azure VM already, So I copy the files onto my VM



1. Then the files must be added to Hadoop





1. Loading and cleaning data with pig.
   1. Create the file tmdb\_clean\_text.pig



* 1. Loading the CSV

A screenshot of a computer

AI-generated content may be incorrect.

* 1. Remove Rows with Missing 'id' or 'name'

A grey background with white text

AI-generated content may be incorrect.

* 1. Remove Duplicates

A white text on a gray background

AI-generated content may be incorrect.

* 1. Remove Invalid or Null 'first\_air\_date'

A white text on a gray background

AI-generated content may be incorrect.

* 1. Normalize 'adult' Column (Convert to Boolean 'true' or 'false')

A screenshot of a computer program

AI-generated content may be incorrect.

* 1. Remove Unreadable Characters (Non-ASCII)

To do this, I created a Java UDF “CleanText.java”.

A computer screen shot of white text

AI-generated content may be incorrect.

Then compile and create a JAR.





Then add the lines to the script.

A screenshot of a computer program

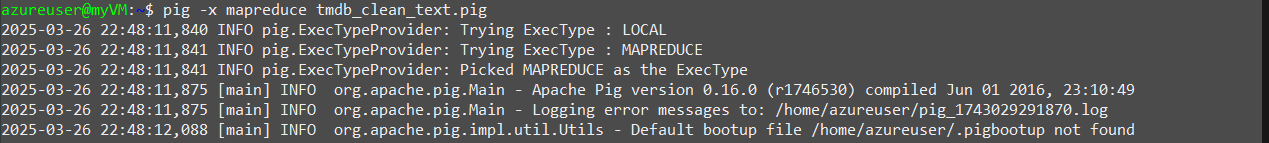
AI-generated content may be incorrect.

* 1. Store the Cleaned Data in HDFS

A close up of a text

AI-generated content may be incorrect.

1. Running the script and Issues



A screenshot of a computer

AI-generated content may be incorrect.

A screen shot of a computer code

AI-generated content may be incorrect.

As you can see from the screenshot, there is only one entry left in the database. I believe this is because of the structure of the original dataset, which has some missing values as well as a column called “languages” that has multiple values per row that throws off how pig expects to read a csv. There was no easy fix that was readily apparent, so we decided to switch our approach to using MongoDB and python.