* Advanced option
* SSH
* hadoop fs -mkdir /user/hadoop/videogames/
* hadoop distcp s3://amazon-reviews-pds/tsv/amazon\_reviews\_us\_Home\_v1\_00.tsv.gz /user/hadoop/amazon\_reviews\_us\_Home\_v1\_00.tsv.gz
* df = spark.read.csv('s3://nyc-tlc/misc/uber\_nyc\_data.csv', header=True)
* df.printSchema()
* df.count()
* df.createOrReplaceTempView("google\_ngrams")
* df1 = spark.sql("SELECT \* FROM google\_ngrams WHERE Token=’Data’").show(10)
* df1.count()
* df1.write.csv(/user/hadoop/eng\_data\_1gram/filter\_data\_1gram)
* hadoop fs -ls
* hadoop fs -getmerge /user/hadoop/eng\_data\_1gram eng\_data\_1gram.csv
* aws s3 cp /user/hadoop/eng\_data\_1gram eng\_data\_1gram.csv s3://brainstation-dsft/eng\_1M\_1gram.csv