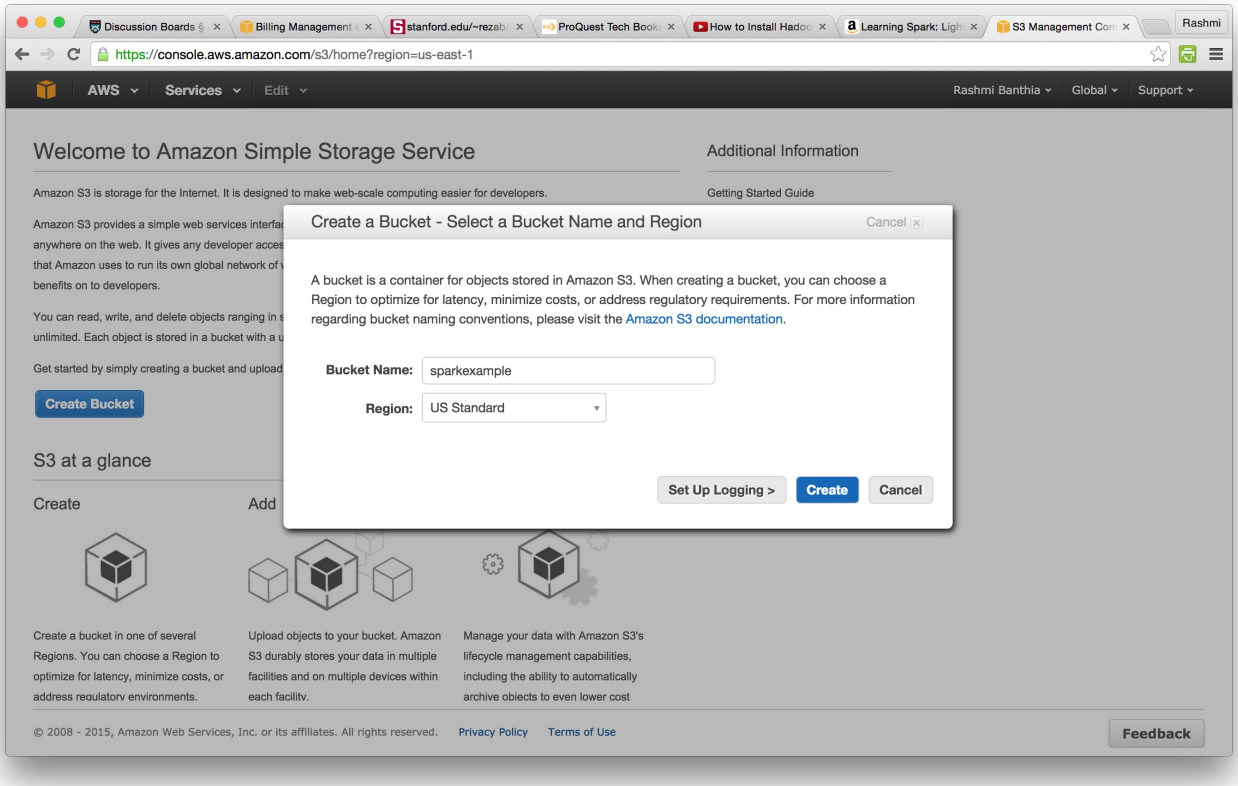
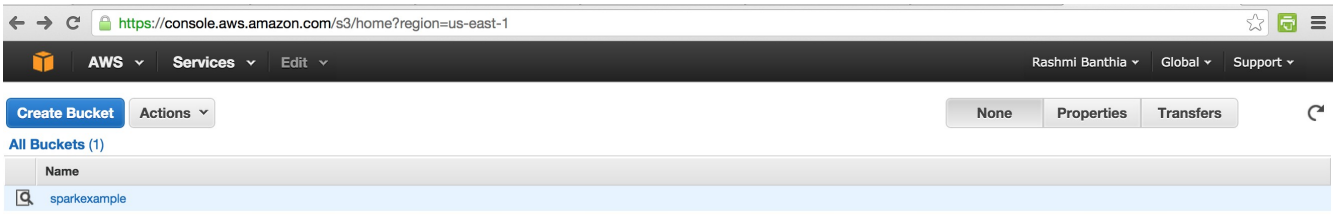
Apache Spark Examples

Creating RDD using S3 bucket

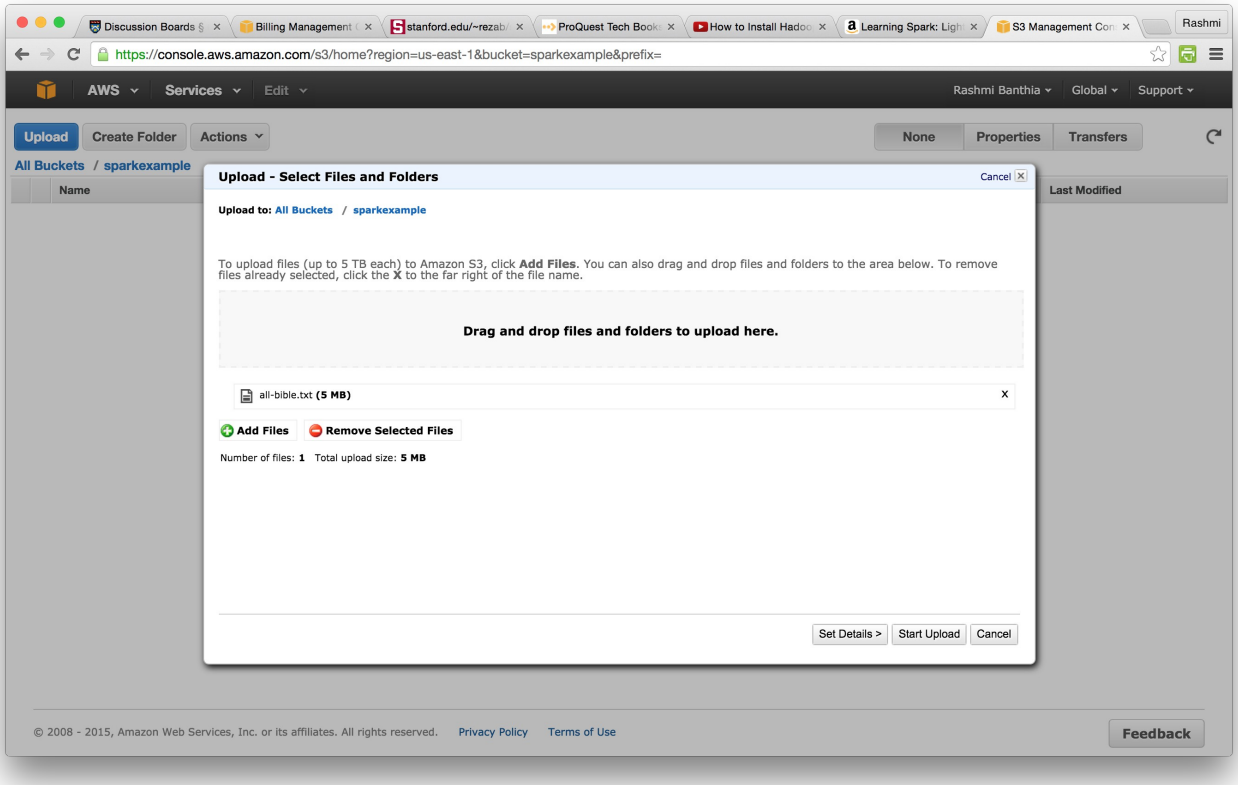
Create a S3 bucket

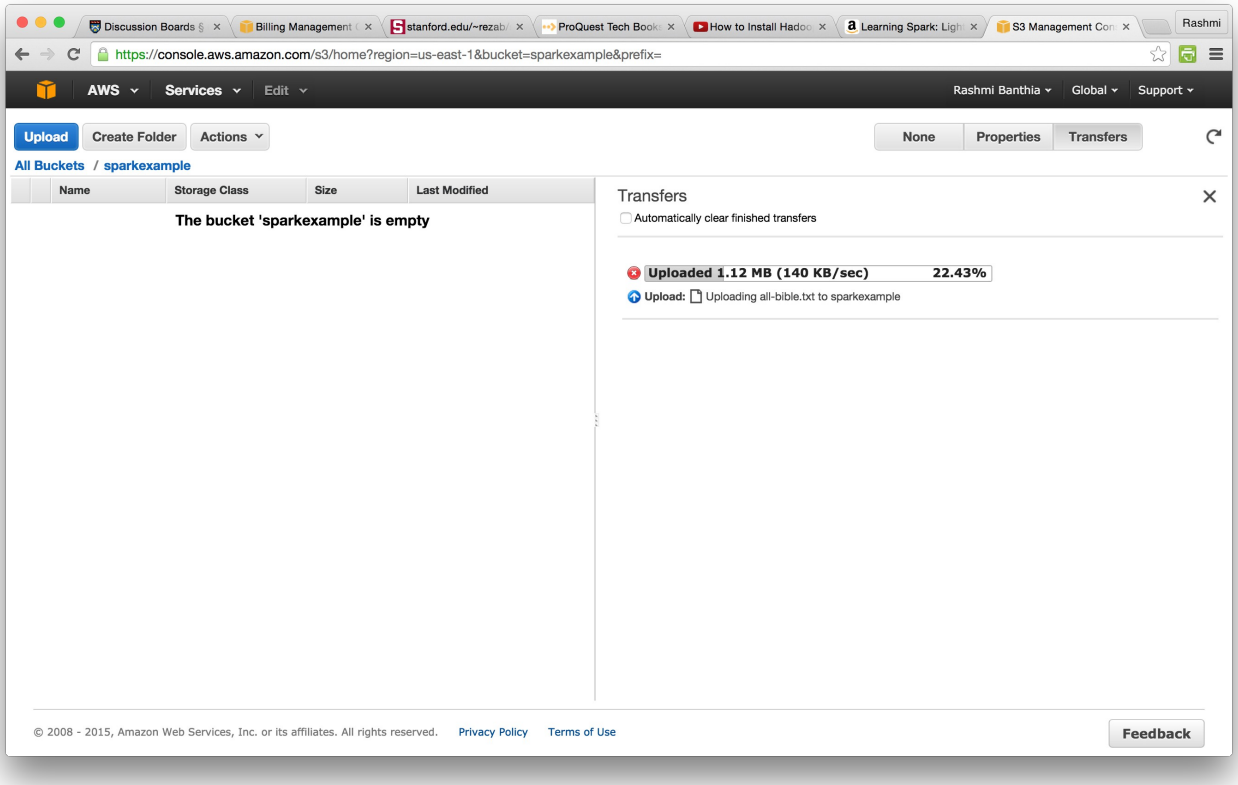


Click on sparkexample



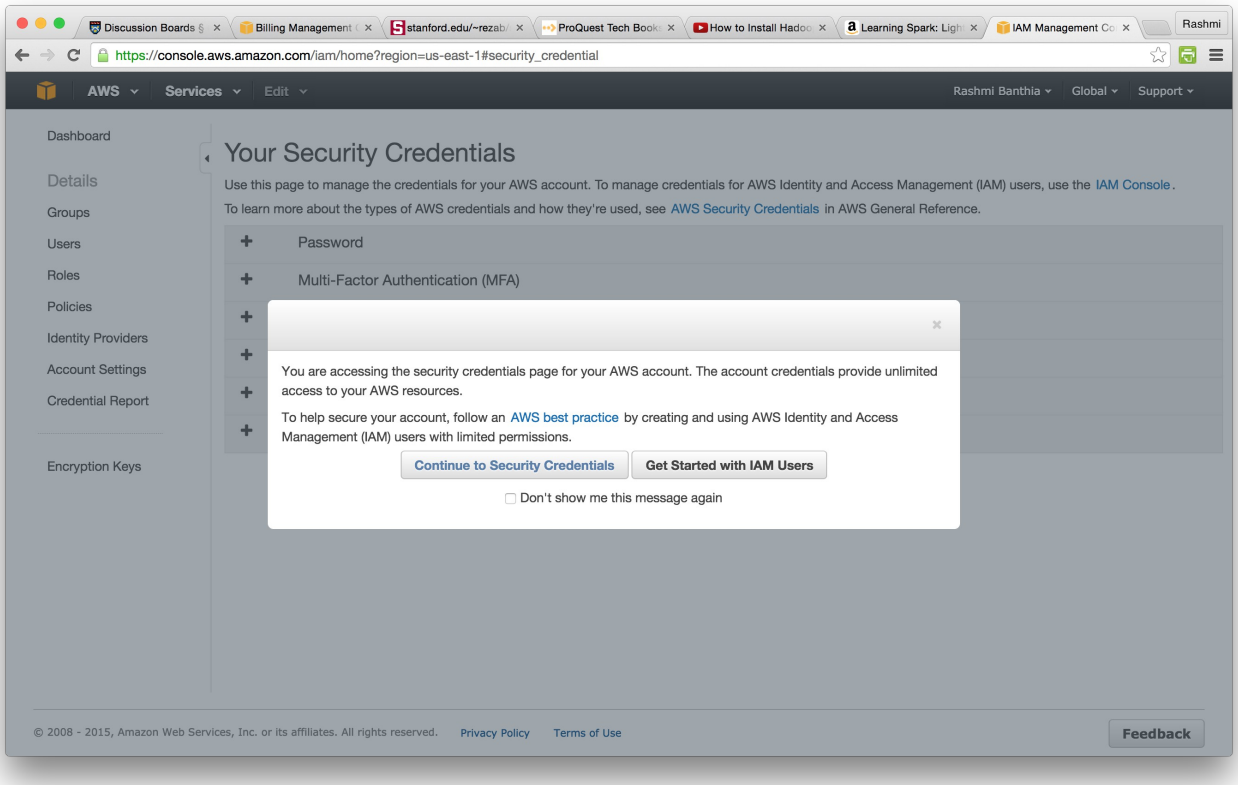
Click on Upload



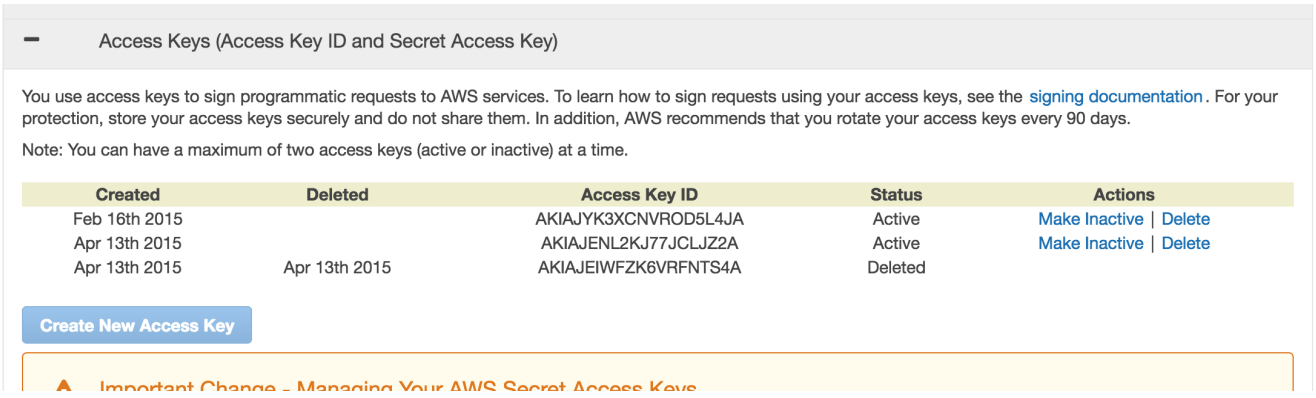


URL of the file:
https://s3.amazonaws.com/sparkexample/all-bible.txt

Click on Security Credentials (top right corner)



Create new Access keys and download it.



Rootkey.csv
AWSAccessKeyId=AKIAJENL2KJ77JCLJZ2A
AWSSecretKey=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

On VM, (or you can put it in .bash_profile)

export AWS_ACCESS_KEY_ID=AKIAJENL2KJ77JCLJZ2A
export AWS_SECRET_ACCESS_KEY=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

(On python spark shell)

$pyspark

>>> s3lines = sc.textFile("s3n://sparkexample/all-bible.txt")
>>> s3lines.count()
117154

Stop/Delete the S3 bucket