**Do not use AWS Services, download a file to your local environment.**

**AWS services are not free!**

Deequ tutorial: <https://aws.amazon.com/ru/blogs/big-data/test-data-quality-at-scale-with-deequ/>

**Task:**

1. Download file from AWS: s3://amazon-reviews-pds/tsv/amazon\_reviews\_us\_Camera\_v1\_00.tsv.gz

Use [AWS CLI](https://docs.amazonaws.cn/en_us/cli/latest/userguide/install-windows.html) (in case of installed python the easiest way is to install via pip).

AWS CLI command:

*aws s3 cp --no-sign-request s3://amazon-reviews-pds/tsv/amazon\_reviews\_us\_Camera\_v1\_00.tsv.gz .*

1. Check that column *verified\_purchase* contains only “N” and “Y” as values
2. Check that column *review\_date* contains only dates (use regex)
3. Check that column *review\_id* contains unique not null values
4. Check that column *total\_votes* contains only integers
5. In case of failed checks investigate the possible reasons with spark.

Link to spark options: <https://spark.apache.org/docs/latest/api/java/org/apache/spark/sql/DataFrameReader.html#csv-scala.collection.Seq->

**As a result**:

Auto: a project with Deequ checks should be provided

Manual: screenshots from spark-shell (scala or python) should be provided