**Overview:**

Download data from link [https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon\_reviews\_us\_Camera\_v1\_00.tsv.gzLinks to an external site.](https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Camera_v1_00.tsv.gz). The description of the columns can be found at https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt (Links to an external site.) (Links to an external site.) .

For this assignment,

* you must run different queries on Apache PIGusing this dataset.
* you must submit all your queries along with the screenshot/files of the results where "Dump" is mentioned.
* every part is dependent on the data insertion in 'part 1'.

**Point Distribution:**

Part 1:

1. Download and Insert this dataset to the PIG server:
2. For each ‘marketplace’ & ‘product category’ combined, find the total number of ‘review\_id’ and average ‘star\_rating’. Dump the result.  
    2.1. Exclude ‘marketplace’ with the value ‘US’ from the previous query. Dump the result.

Part 2:

1. For each ‘review\_date’, find the total number of ‘review\_id’. Dump The Result
2. For each ‘product\_id’, find the average ‘helpful\_votes’ and average ‘total\_votes’. Dump the result.  
    2.1. Show top 10 results of the previous query based on the  average ‘total\_votes’. Dump the result.

Part 3:

1. Consider only the ‘verified purchase’ (i.e. Filter out unverified purchase) for the following queries.

            1.1. For each ’product category’, find the total number of  ‘product\_id’. Dump the result.

            1.2. For each ‘star\_rating’, find sum of ‘helpful\_votes’, sum of ‘total\_votes’, and avg\_votes (i.e sum of ‘helpful\_votes’ divided by sum of ‘total\_votes’). Dump the result.

                        1.2.1. Apply ‘Order by avg\_vote’ on the previous query. Dump the result.