# Big Data Hadoop Project

## Problem Statement:

Imagine you work for an e-commerce website as a Big Data analyst. Millions of people put reviews on the products on your website. You have been tasked with running an analysis on these reviews.

## Datasets (amazondataset and wordsentimentdataset):

https://intellipaat-course-attachments.s3.ap-south-1.amazonaws.com/Hadoop/hadoop_dataset.rar

## Tasks To Be Performed:

1. Analyze all the positive reviews (any review with a rating of 4 and above is considered positive) and find out the top 20 words used in those positive reviews.
2. Use the word sentiment dataset and find out the percentage of words that are positive, negative and neutral. The words that aren't mentioned in the word sentiment dataset are considered as neutral.

**Note:** The attributes, 'reviewsrating' and 'reviewstext' correspond with the ratings and the text reviews, respectively.

**Note:** In the word sentiment dataset, a positive numeric value denotes a positive emotion and a negative numeric value denotes a negative emotion attached to the word.

For this project, you are allowed to use any method within the course syllabus.

## Constraints:

- For accuracy, do not treat uppercase words and lowercase words as separate words. Convert all words to lowercase. (For example: 'His' and 'his' should not be treated as separate words)

  If you have a sentence "His his", then the output should be:

  his 2

 

And not:

his 1
His 1

- Remove all punctuation marks from the beginning and the end of all words: ‘hello?’, ‘hello.’, ‘hello-‘, ‘hello’, ‘HellO’

  will all correspond to the following count:

  hello 5