# <u>Project 1 Phase 3 – Hadoop, MapReduce,</u> Spark, Fasttext

**Due: 6th 11:59pm** 

### **Learning Outcomes for Phase 3:**

- 1. Create local or remote instances of Hadoop and load your dataset 5 marks
- 2. Identify the feature you want to analyze and perform word count or any other task using spark context or MapReduce. 10 marks
- 3. Explain the working of MapReduce on the data you selected for task2. 5 marks

#### **Deliverables:**

- 1. Loading data in Hadoop [5 Marks]:
- Create local or remote instance of Hadoop.
- Load all the data that you collected doesn't have to be just 2000 rows.
- Load all the data that you collected into a pandas dataframe.
- Time comparison between Hadoop and dataframe loading.
- 2. Using MapReduce [10 marks]:
- Select a feature from your dataset that you want to analyze using MapReduce.
- Perform word count or any other task of your choice on all the data using MapReduce or Sparkcontext (not just 2000 rows).
- Perform above step without MapReduce.
- Time comparison for above 2 steps.
- 3. Working of MapReduce [5 marks]:
- Similar to the flow chart we discussed in lectures.

#### Shuffling Splitting Mapping Reducing Final Result Input List(K2, V2) K2, List(V2) K1, V1 Deer, 1 Bear, 1 River, 1 List(K3, V3) Bear, (1,1) Deer Bear River Car, (1,1,1) Bear, 2 Car, 3 Deer, 2 River, 2 Deer Bear River Car Car River Deer Car Bear Car, 1 Car, 1 River, 1

The Overall MapReduce Word Count Process

edureka!

MapReduce Example - MapReduce Tutorial

Deer, Car, Bear,

Deer, (1,1)

River, (1,1)

River, 2

## **Extra Credit:**

- 1. Word2Vec/Fasttext [10 marks]:
- Use word2vec or Fasttext to perform sentiment analysis or similarity search on textual data from your dataset. (limit to 2000 rows)

Car Car River

Deer Car Bear