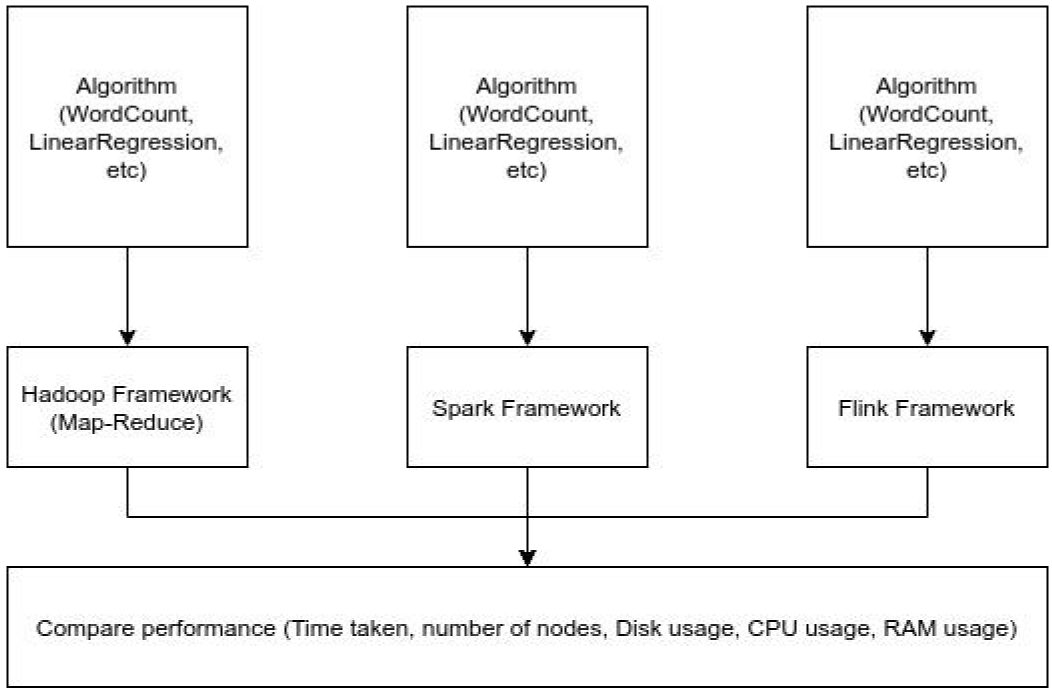
Evaluation of Big Data frameworks using different Machine Learning algorithms

Methodology



Big Data frameworks:

The Big Data frameworks that will be considered for evaluation are,

1. Hadoop
2. Spark
3. Flink

Algorithms to be evaluated:

The following algorithms that will be evaluated using the 3 Big Data frameworks:

1. Word Count
   Word count is an algorithm used to get the count of words in an input text i.e the number of occurrences of a word in the file. This algorithm will be used as a reference algorithm.

2. Linear regression

   Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). This is used to predict continuous values given a set of inputs.

3. Decision Tree
   Decision tree learning is one of the predictive modeling approaches used in statistics, data mining and machine learning. It uses a decision tree to go from observations about an item to conclusions about the item's target value.

Hardware used for evaluation:

CPU: Intel® Core™ i3-8130U CPU @ 2.20GHz × 4

Memory: 11.6 GiB

Disk: 89.1 GB

OS: Ubuntu 18.04 (Linux)

Parameters to be evaluated:

The following parameters will be compared

1. Disk Usage

   It is the amount of disk space used by the framework. It is better to have the disk usage low.

   The following command will be used to get the disk usage for each of the frameworks

   du -sh <path to framework>

2. RAM Usage

It is the amount of memory used by the framework when an algorithm is executing. It is better to have this value low.

The following command will be used to get the RAM usage for each of the frameworks

top | grep <framework process name>

3. CPU Usage

   This will give the load on the CPU when an algorithm is executing. It is better to have this value low.

   The following command will be used to get the CPU usage for each of the frameworks

   top | grep <framework process name>

4. Execution time vs Input data size

   Input dataset size will be varied and the execution time will be measured in each case.

5. Execution time vs Number of nodes(workers)

   Number of nodes in the cluster will be varied and the execution time will be measured in each case.

6. Execution time (different algorithm)