Proposal Document

- Project Overview

  Analysis of machine learning algorithm performance with distributed computing systems

- Problem Definition

  The processes involved in Machine Learning (training/testing) are very computationally intensive and take a lot of time for completion. With lots of computational systems readily available, our goal is to analyse the performance of ML algorithms when they are parallelized.

- Scope

  This project will be useful for people having multiple mid-performance computers and have to process large amount of data and apply machine learning techniques to solve problems

- Objectives
  - Parallelize Machine Learning algorithm process(training/testing)
  - Document the time taken with many computational systems starting with one
  - Find the optimum number of computational systems needed for best performance.

- Generic project life cycle for the chosen technology

- Build a distributed system capable of solving simple problems parallelly.
- Identify a CPU intensive Machine Learning algorithm
- Parallelize the ML process and solve it with the help of the distributed system built initially.
- Increase/decrease the number of computational units and find the optimum number of computational systems.

- Planned Technology to be used
  - Programming language: Python
  - Machine Learning framework: PyTorch, Tensorflow, scikit-learn
  - Distributed Systems/Parallel processing framework: Celery, Ray, Parallel Python, RabbitMQ