Progress Report

1. Principal Investigator

Matthew DiDomizio (didomizm1@newpaltz.edu)
Maranda Dominguez (domingum6@newpaltz.edu)

2. Title of Project

Anime Recommendation System

3. Project Progress Summary

So far, we have set up the Hadoop cluster, dealt with our dataset, and begun programming our MapReduce algorithm. A GitHub repository was created and structured, allowing for collaboration on the project, and it holds both the data and code which will be ready for use with Hadoop. After sourcing our data, we went through a data analysis stage to determine which algorithm would best fit our vision for the recommendation system and we made decisions as to what parts of the data were needed and not needed. Subsequently, we cleaned the data using Pandas and formatted it properly for our use, also generating statistics on how the data had changed after our alterations. Once the data was successfully cleaned and prepared for use, we begun to program our map and reduce functions in Python using the K-Means clustering machine learning algorithm.

4. Activities

1. 9/30/22: Set up the Hadoop cluster and created a GitHub repository with a clearly defined file structure in order to hold the project data and code, as well as to allow for collaboration and version control.
2. 10/7/22: Discussed the matter of which dataset to use to carry out the vision for our proposed project.
3. 10/14/22: Discussed the analysis of the dataset, ensuring its proper formatting, and cleaning the data in accordance with the parts of the data that were relevant to our chosen machine learning algorithm.

4. 10/21/22: Considered the programming of the K-Means algorithm in terms of MapReduce and discussed its implementation using Python.
5. 10/28/22: Continued our progress with the core algorithm programming and discussed, as well as completed, the progress report.

5. Remaining Work

As we progress further into the project, there remains quite a bit left to do. Now that we have collected and cleaned our data, we have yet to finish the programming of our map and reduce functions. These functions will be utilized by Hadoop and will be implemented using the K-Means clustering algorithm. Once we have completed the map and reduce functions, we must perform machine learning and test our algorithm in order to confirm that it is functioning properly. When we have successfully verified the functionality of our algorithm and are assured that it is working as intended, we can proceed with visualizing our data along with our findings. The Python libraries Matplotlib and Seaborn will be used in order to accomplish the necessary data visualization needed to present our findings and further analyze the data.

6. Difficulties

Considering the large size of our dataset, we came to the conclusion that we would need to find the best possible way to sift through the data, analyze its contents, and decide which parts were needed, unneeded, and improperly formatted for our intended use. After much discussion and deliberation, we decided upon the use of the Python library, Pandas, to meet our goals on this front, and it ultimately turned out to be the correct choice to deal with this data at scale. In addition, the dataset itself was a matter of contention; indubitably, it was of the utmost import that our data reflected the requirements of our intended objectives for this project, and as such, the choice as to which set of data to serve as a foundation for our recommendation system was carefully thought out. In particular, it was crucial that the data itself would be compatible with the K-Means machine learning algorithm, as well as that the data would be large enough to warrant accurate predictions as to a future user's likely series preferences. In the end, we successfully found a dataset which suited our needs. As a final point, the decision to make use of Python was not without its own tribulations – indeed, it can be said that both Java and Python presented us with their own distinct and greatly alluring qualities, in essence beckoning us to either side of the proverbial aisle. While it is the case that

Java is, perchance, the better integrated of the two with Hadoop, Python’s lightweight disposition and proprietorship of useful data science libraries made the subsequent choice all but too clear. Python was undoubtedly the more prudent option in allowing us to work towards the forthcoming completion of our novel innovation.

7. Significant Changes to Proposal (if any)

N/A