Task 2 (write a MapReduce program for Hadoop in Python)

Choose a small dataset, analyse it to get some meaningful result. You have to clarify your dataset format and what you will get from it. Example:

Given two datasets:

- A set of user demographic information containing [id, email, language, location]
- A set of item purchases, containing fields [transaction-id, product-id, user-id, purchase-amount, product-description]

Calculate the number of locations in which a product is purchased.

Task 3 (nothing new)

What you will submit:

When you have finished implementing the complete your 2 tasks as described above, you should submit your solution on drive link on the course page:

1. **Output file:** A copy of the output generated by running your tasks. Ex: When it downloads a file, have your program print a message "display file 'foo'" (don't print the actual file contents if they are large). When a peer issues a query (lookup) to the indexing server, having your program print the returned results in a nicely formatted manner.
2. **Design Doc:** A separate (typed) design document (named design.pdf or design.txt) of approximately 2-4 pages describing the overall tasks design, and design trade-offs considered and made. Also describe possible improvements and extensions to your program (and sketch how they might be made).
3. **Manual:** A detailed manual describing how the tasks work. The manual should be able to instruct users other than the developer to run the tasks step by step. The manual should contain at least one test case which will generate the output matching the content of the output file you provided in 1.
4. **Verification:** A separate description (named test.pdf or test.txt) of the tests you ran on your program to convince yourself that it is indeed correct. Also describe any cases for which your program is known not to work correctly.
5. **Source code:** All of the source code for all tasks.