Map Reduce Report

- pnalawa@iu.edu

Design:

- Map Reduce is implemented in these many parts:
  - Master
  - Mapper
  - Mapper Task
  - Reducer
  - Reducer Task
  - Library

- **Master**
  - Acts as coordinator of all the processes
  - Flask app is exposed at port 8000
  - Starts Mapper server at port 8002 // can be changed by .env or exporting MAPPER_PORT
  - Starts Reducer server at port 8003 // can be changed by .env or exporting REDUCER_PORT
  - Spawns one server process
  - Server
    - This is used to create mapper server and it spawns N processes for N mappers
    - This manages the mapper barrier part which waits for all mapper tasks to complete
    - This manages shuffling and sorting.
    - This is used to create reducer server and it spawns N processes for N reducers
    - This manages the reducer barrier part which waits for all reducer tasks to complete
  - Spawns N Mappers clients to connect to mapper task
    - These N mapper clients connect to server mapper server which is accepting connections.
  - Spawns N reducers clients to connect to reducer task
    - These N mapper clients connect to server mapper server which is accepting connections.

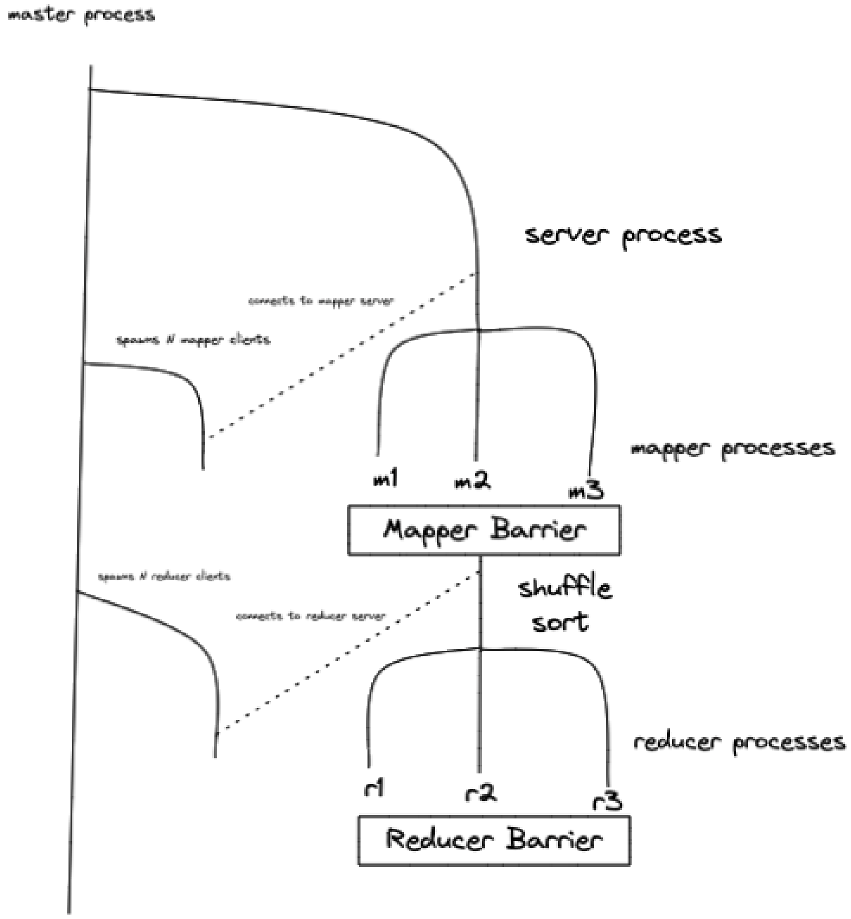
- **Mapper**
  - It handles cleaning of data and applying mapper operations and saving all the post processed data in the dictionary.

- **Reducer**
    - It handles cleaning of data and applying reducer operations and saving all the post processed data in the dictionary.
- **Mapper Task**
    - It handles sending and receiving mapper data through sockets.

- **Reducer Task**
    - It handles sending and receiving mapper data through sockets.

Architecture

Note: The below diagram shows for 3 mapper and 3 reducer workers but ideally it **works for N mappers and N reducers.**



- **Data Partitioning**
    - **Word Count**

    - Suppose, we have one book of data.
    - Divides length of data by the number of mappers.
      - Eg: len(data) / number_of_mappers
    - For reducers send, the KV of the mappers output based on hashing of all the characters of key modulus number of reducers
      - Eg:hash(key) % number_of _reducers

  - **Inverted Index**
    - Suppose we have number of books
    - Send books to mappers by modulus of mappers. At Least one book is send to each mapper if number of books is greater than number of mappers
      - Eg: index_of_book % number_of_mappers
    - Send KV by hashing by the formula below:
      - hash(key) % number_of _reducers

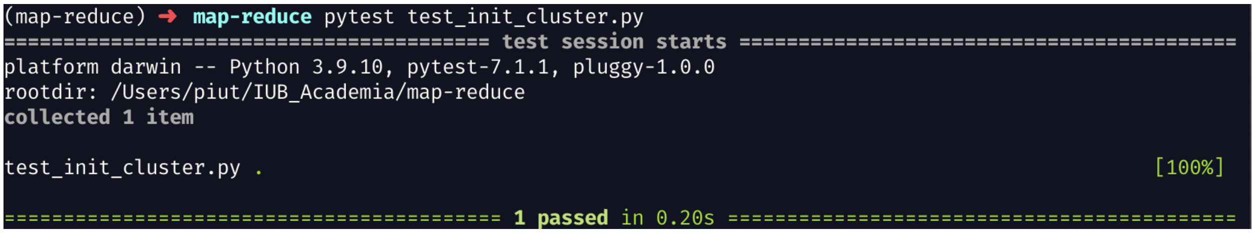
- **Message Format**
  - **Header data** of the mapper and reducer are being sent to the mapper task by **protobuf**
    - Header data contents length of data to be sent.
  - **Dictionary data / Raw data** is sent to the mapper task and reducer task by **pickling which is bytes.**
    - This contains actual data to be sent.

- **Library**
  - Two functions exists in this Map Reduce Library class
    - **Init_cluster**
      - **Parameters are:**
        - *"""*
        - *Initiates Map Reduce Cluster*
        - *:param map_count: integer*
        - *:param count_reducer: integer*
        - *:return: string*
        - *"""*
        -
    - **Run_mapred**
      - **Parameters are:**
        - *"""*
        - *Starts map reduce*
        - *:param input_file: string*
        - *:param operation_type1: string (word_count, inverted_index)*
        - *:param operation_type2: string (word_count, inverted_index)*
        - *:param output_file: string*
        - *:return: string*

- Library functions make API calls to the flask server which is running on master at port 8000.

- **Steps to Run**:
    - cd map-reduce
    - brew install pipenv
    - export MAPPER_PORT = <any valid port>
    - export REDUCER_PORT = <any valid port>
    - pipenv shell # initiates virtual environment
    - pip install -r requirements.txt
    - python master.py # creates master and flask app on 8000
    - pytest test_init_cluster.py -v
    - pytest test_wc.py -v
    - Close python master.py and run again
    - pytest test_ii.py

- **Test Cases Outputs & Screenshots:**
    - Init Cluster

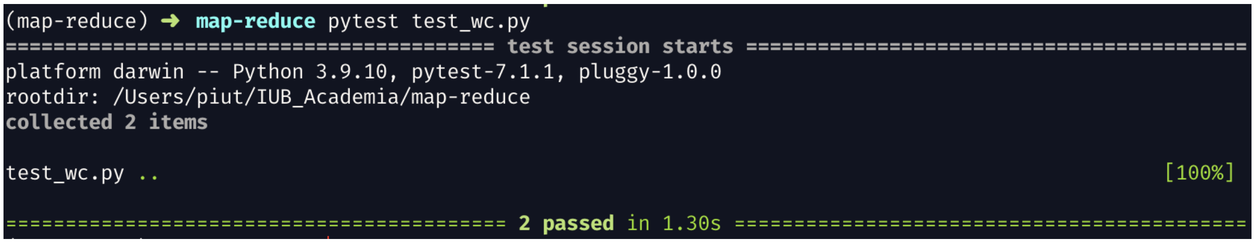


```
(map-reduce) ➜ map-reduce pytest test_init_cluster.py
======================================== test session starts =========================================
platform darwin -- Python 3.9.10, pytest-7.1.1, pluggy-1.0.0
rootdir: /Users/piut/IUB_Academia/map-reduce
collected 1 item

test_init_cluster.py .                                                                          [100%]

========================================= 1 passed in 0.20s ==========================================
```

- Word Count

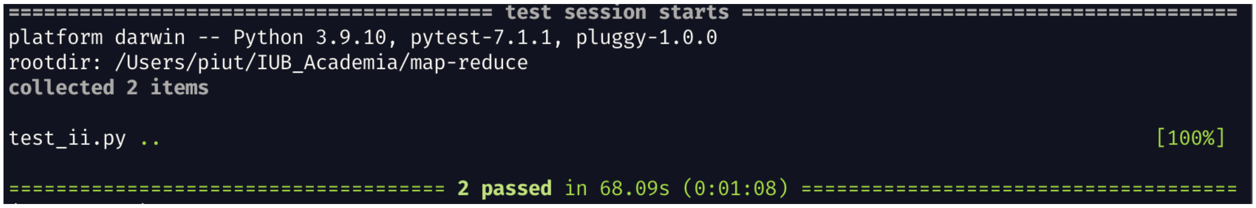


```
(map-reduce) ➜ map-reduce pytest test_wc.py
======================================== test session starts =========================================
platform darwin -- Python 3.9.10, pytest-7.1.1, pluggy-1.0.0
rootdir: /Users/piut/IUB_Academia/map-reduce
collected 2 items

test_wc.py ..                                                                                   [100%]

========================================= 2 passed in 1.30s ==========================================
```

- Inverted Index



```
========================================= test session starts ==========================================
platform darwin -- Python 3.9.10, pytest-7.1.1, pluggy-1.0.0
rootdir: /Users/piut/IUB_Academia/map-reduce
collected 2 items

test_ii.py ..                                                                                    [100%]

===================================== 2 passed in 68.09s (0:01:08) =====================================
```

Output Word Count:

You can see output in **word_count/results.txt**

Output Inverted Index:

You can see output in **inverted_index/results.txt**