과목명: 시스템프로그래밍

담당교수명 : 소정민

<<Assignment 6>>

서강대학교 컴퓨터공학과

이정원20171672

목차

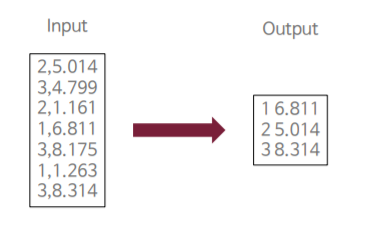
1. **프로그램 개요3**
2. **프로그램 설명3**
3. 모듈 정의**3**
   1. 3
   2. 3
4. **코드4**
5. **첨부 사진5**
6. **프로그램 개요**

이 프로젝트에서는 Hadoop과 MapReduce의 기초에 대해 이해하는 것을 목적으로 한다.

1. **프로그램 설명**

* **Group by max**

정수와 콤마, 그리고 실수로 이루어진 파일을 읽어들여 정수 부분의 그룹에서 가장 값이 큰 값을 그룹별로 출력하기 위한 문제이다.



input.data 를 generate.py를 이용하여 16개의 그룹으로 나뉘어진 1억개의 실수로 만들고, Hadoop MapReduce를 이용하여 결과를 출력한다.

1. **모듈 정의**
   1. **mapper.py**

|  |  |
| --- | --- |
| 풀이 및 알고리즘 | 파일을 읽어 들여와 ‘콤마’로 group (정수 부분)과 number (실수 부분)을 split()함수를 이용해 분리하여 mapping하고, 그것을 출력한다. |

* 1. **reducer.py**

|  |  |
| --- | --- |
| 풀이 및 알고리즘 | mapper.py에서 출력된 것을 파일로 읽어들여 콤마로 구분하여 group과 number을 분리하고, number은 문자열이므로 실수화를 시켜 준다. data라는 딕셔너리의 key는 group이 되고, 해당 group에 number의 최대값을 max함수를 이용하여 넣어 준다. 마지막으로 ValueError을 except 시켜 pass 한다.  그리고 결과를 출력한다. |

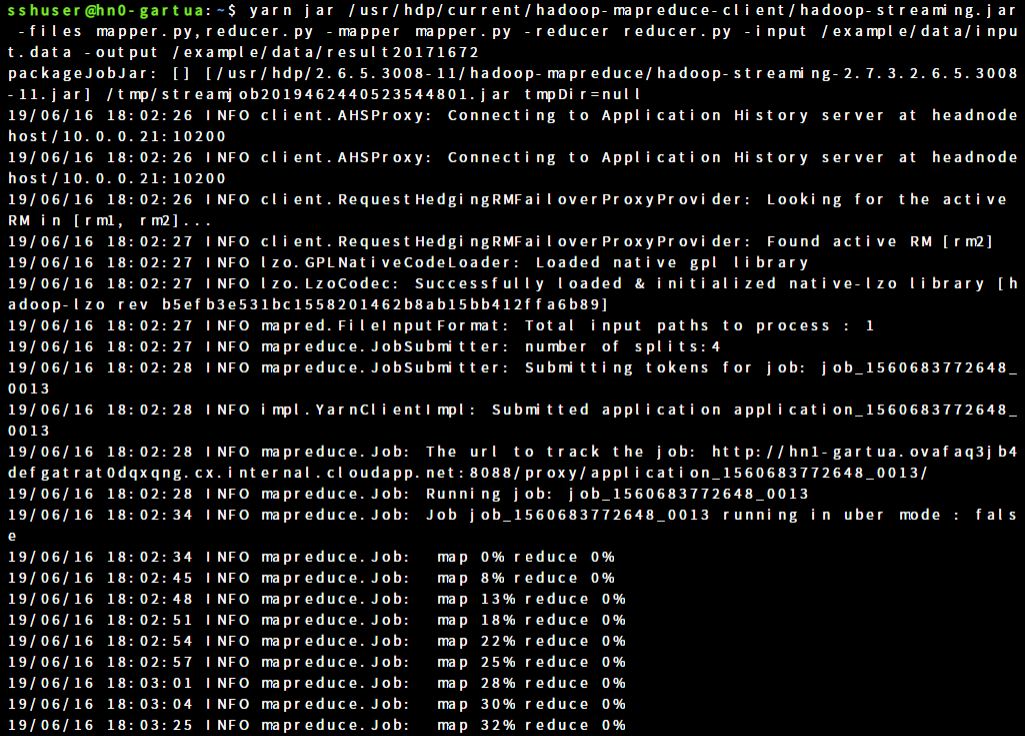
1. **코드**

|  |
| --- |
| **mapper.py** |
| **#!/usr/bin/env python**  **# Use the sys module**  **import sys**  **def main(separator=','):**  **for index, line in enumerate(sys.stdin):**  **group, number = line.split(",")**  **print('%s%s%s' % (group, separator,number))**  **if \_\_name\_\_ == "\_\_main\_\_":**  **main()** |

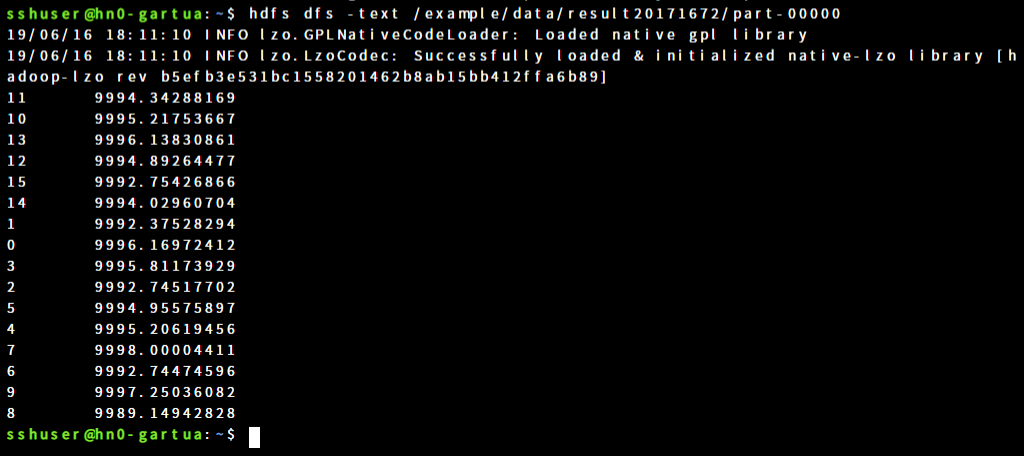
|  |
| --- |
| **redcuer.py** |
| **#!/usr/bin/env python**  **# import modules**  **from itertools import groupby**  **from operator import itemgetter**  **import sys**  **def main(separator='\t'):**  **data = {}**  **for index, line in enumerate(sys.stdin):**  **try:**  **group, number = line.split(",")**  **number = float(number)**  **if group in data.keys():**  **data[group] = max(data[group], number)**  **else:**  **data[group] = number**  **except ValueError:**  **pass**  **for group, number in data.items():**  **print("{} {}".format(group, number))**  **if \_\_name\_\_ == "\_\_main\_\_":**  **main()** |

1. **첨부 사진**

**20171672\_1.png**

****

**20171672\_2.png**

****