Map Reduce

In this assignment we have used map reduce to analyze the university students grades dataset using mrjob library. MapReduce enables parallel computation, improving efficiency, scalability, and fault tolerance when handling large datasets.

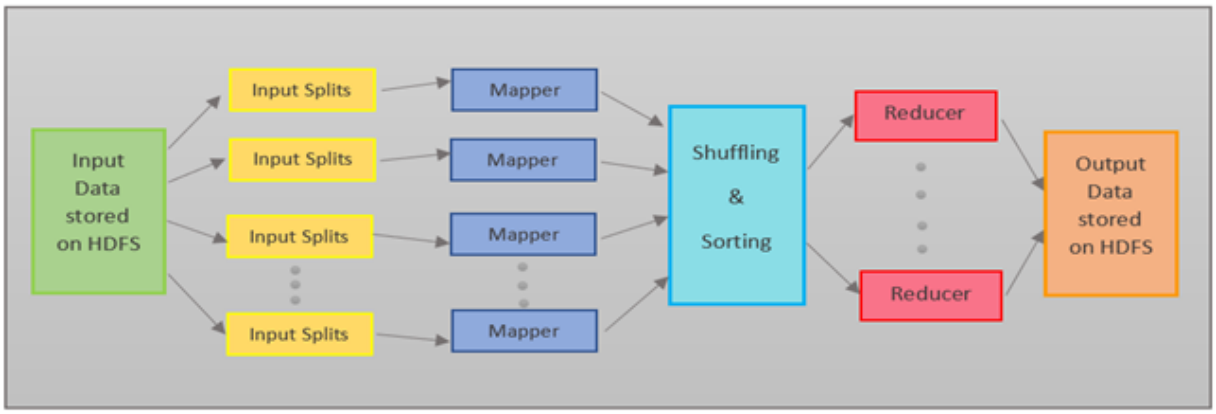
Each task follows the **MapReduce framework**, consisting of:

1. **Input Splitting**: The dataset is divided into smaller parts.

2. **Mapping**: Key-value pairs are generated from each data entry.

3. **Shuffling & Sorting**: Intermediate key-value pairs are grouped.

4. **Reducing**: Final computation is performed to summarize results.



This algorithim eliminates the bottleneck effect from the traditional low level operating system(OS) algorithims like: shortest job first(SJF), first come first serve(FCFS), Round

Robin(RR). As it computes all jobs concurrently with equal slices so it achived the elimination of the bottleneck effect.

MRJOB: MRJob is a Python package that simplifies the process of writing MapReduce jobs. It provides a high-level API that abstracts away the low-level details of Hadoop MapReduce, making it easier for developers to write and run MapReduce jobs.

Task 1: Compute the Average Grade Per Course

Problem:

We need to calculate the **average grade** for each course and determine which course has the highest average.

Mapper Phase:

• Reads each line from the dataset.

• Extracts the Course Name and Grade.

• Returns (Course Name, Grade) pair.

Shuffling & Sorting:

• Groups all grades by Course Name.

Reducer Phase:

• Computes the average grade per course.

• Returns (Course Name, Average Grade).

Task 2: Compute the Average Grade Per University

Problem:

We need to calculate the **average grade** for each university and identify which university has the highest average.

Mapper Phase:

• Reads each line from the dataset.

• Extracts the **University Name** and **Grade**.

• returns (University Name, Grade) pair.

Shuffling & Sorting:

• Groups all grades by University Name.

Reducer Phase:

• Computes the average grade per university.

• Emits (University Name, Average Grade).

Task 3: Identify the Top 3 Highest Grades Per Year

Problem:

Mapper Phase:

• Reads each line from the dataset.

• Extracts the Year and Grade.

• returns (Year, Grade) pair.

Shuffling & Sorting Phase:

• Groups all grades by Year.

• Sorts the grades in descending order.

Reducer Phase:

• Selects the top 3 grades for each year.

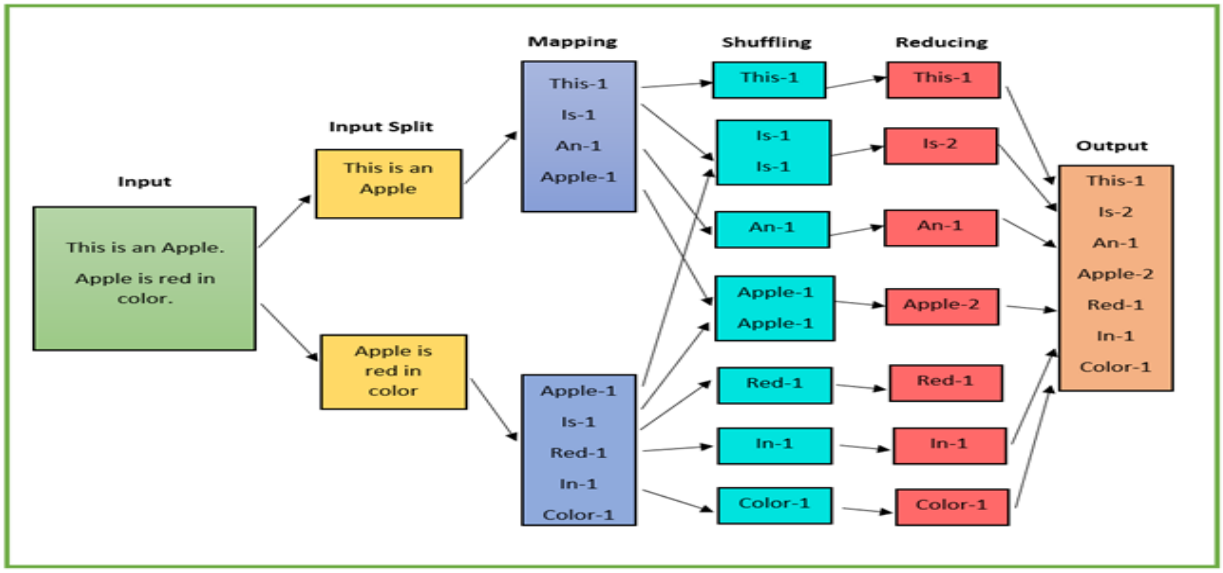
• Returns (Year, [Top 3 Grades]).

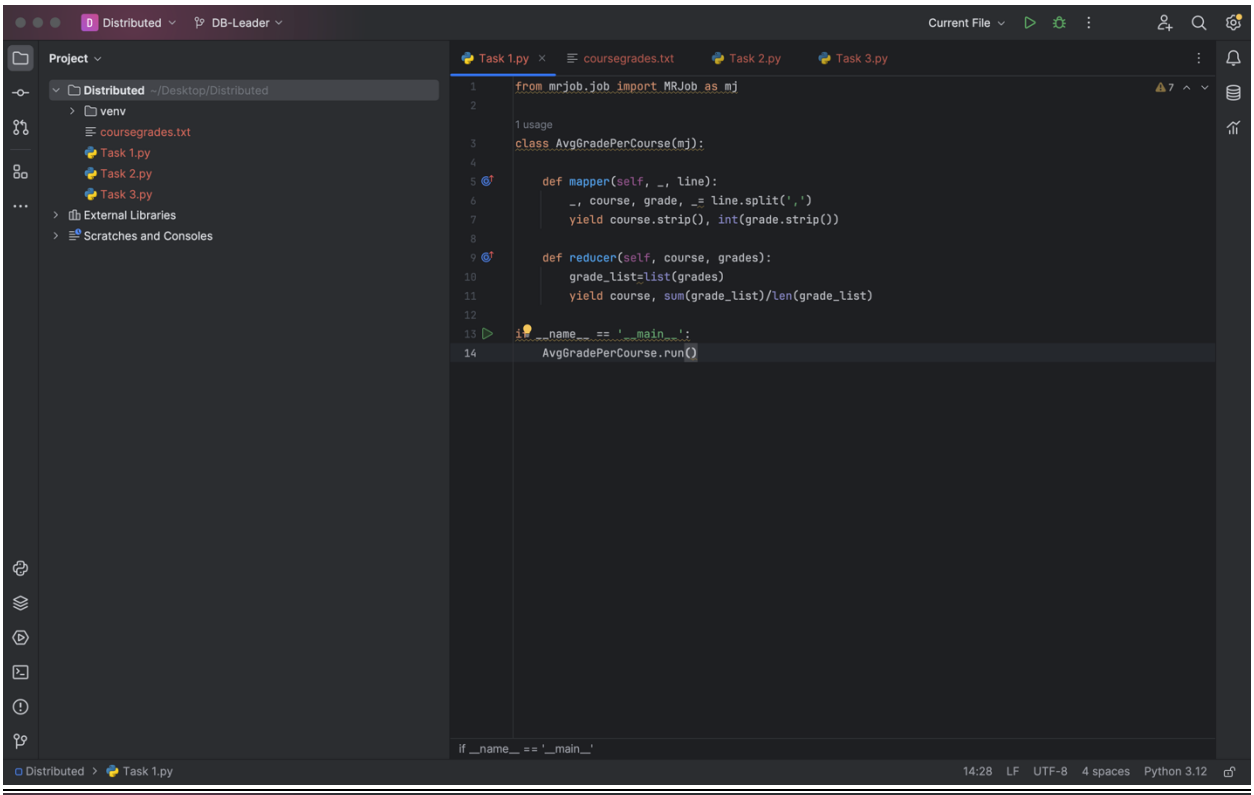
Code Execution:

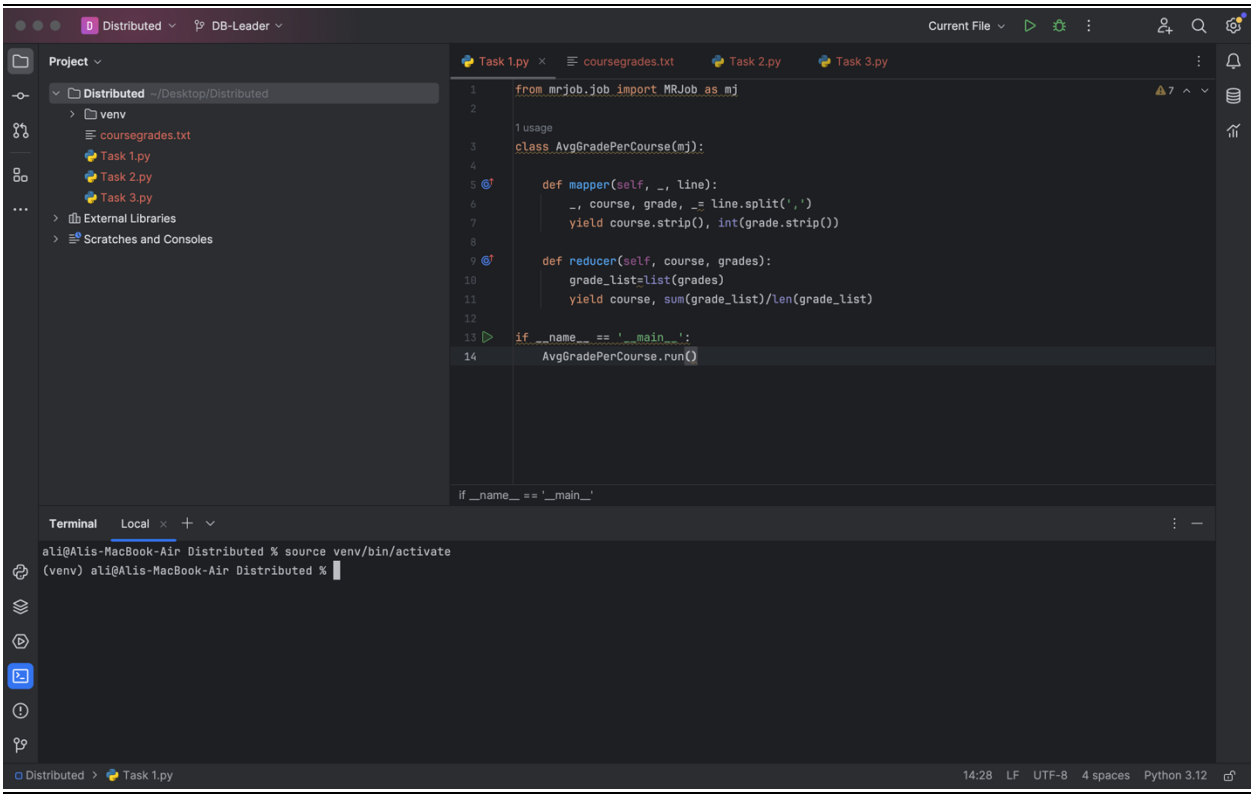
python Task1.py coursegrades.txt > output_task1.txt

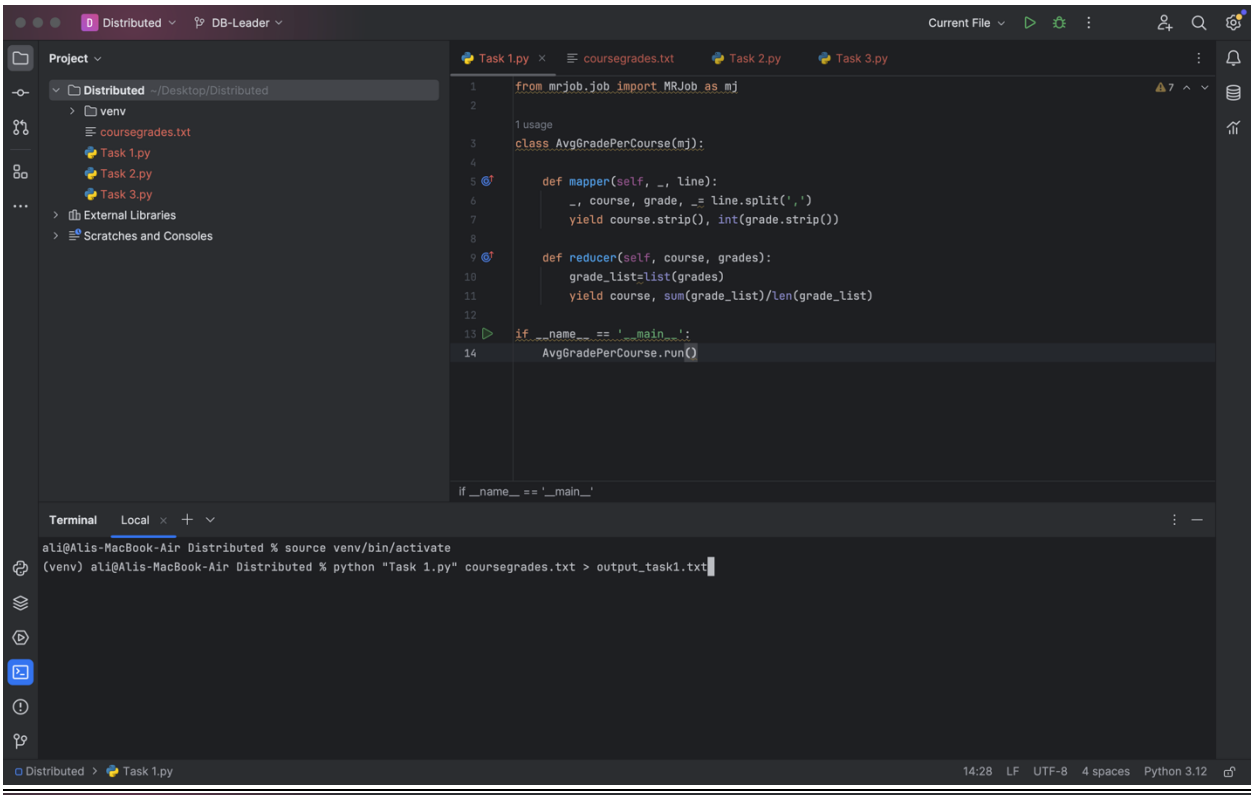
python Task2.py coursegrades.txt > output_task2.txt

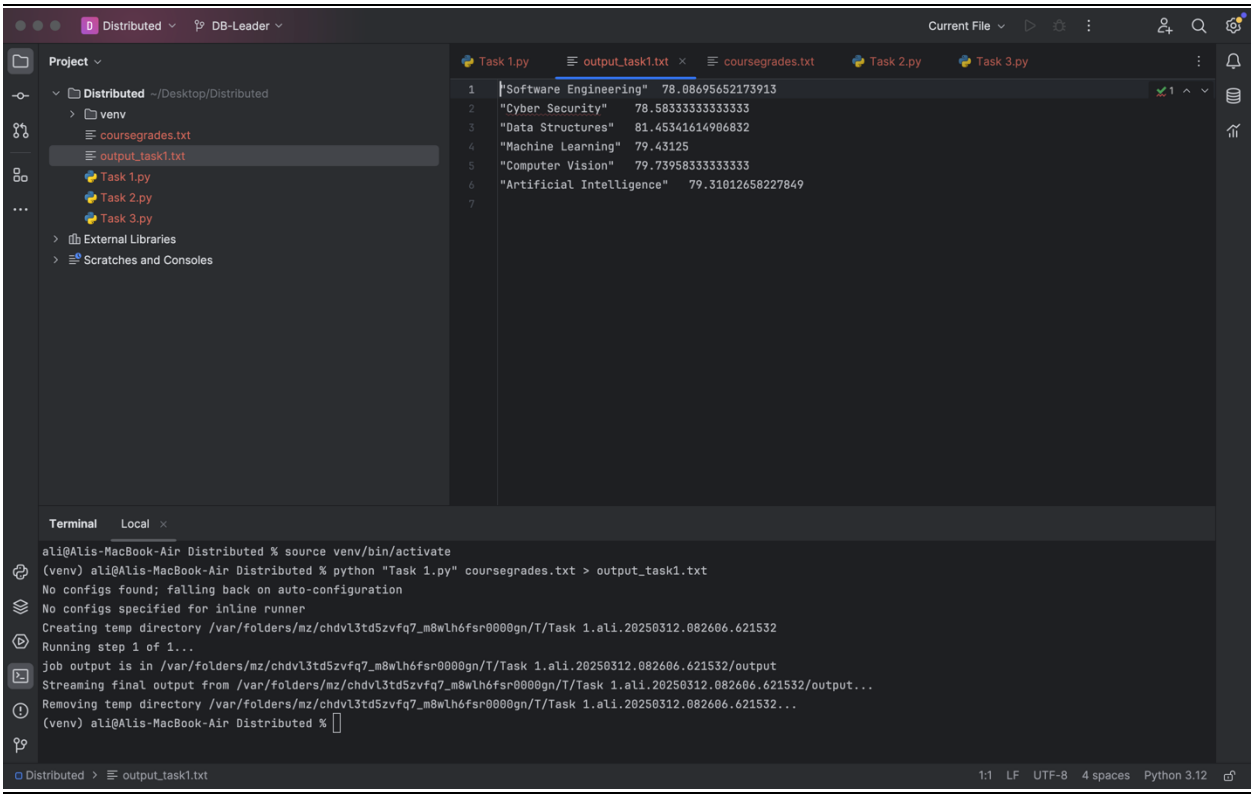
python Task3.py coursegrades.txt > output_task3.txt

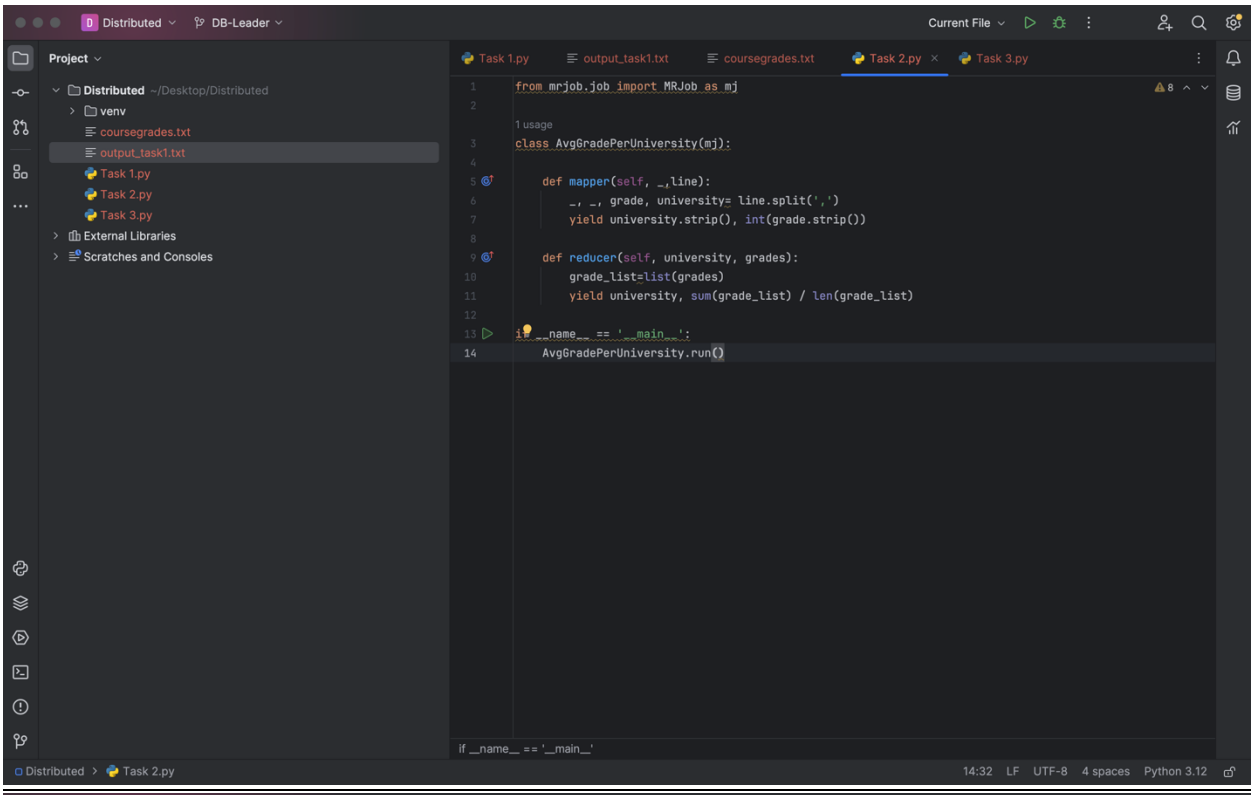

Input
This is an Apple.
Apple is red in color.
Input Split
This is an Apple
Apple is red in color
Mapping
This-1
Is-1
An-1
Apple-1
Apple-1
Is-1
Red-1
In-1
Color-1
Shuffling
This-1
Is-1
Is-1
An-1
Apple-1
Apple-1
Red-1
In-1
Color-1
Reducing
This-1
Is-2
An-1
Apple-2
Red-1
In-1
Color-1
Output
This-1
Is-2
An-1
Apple-2
Red-1
In-1
Color-1

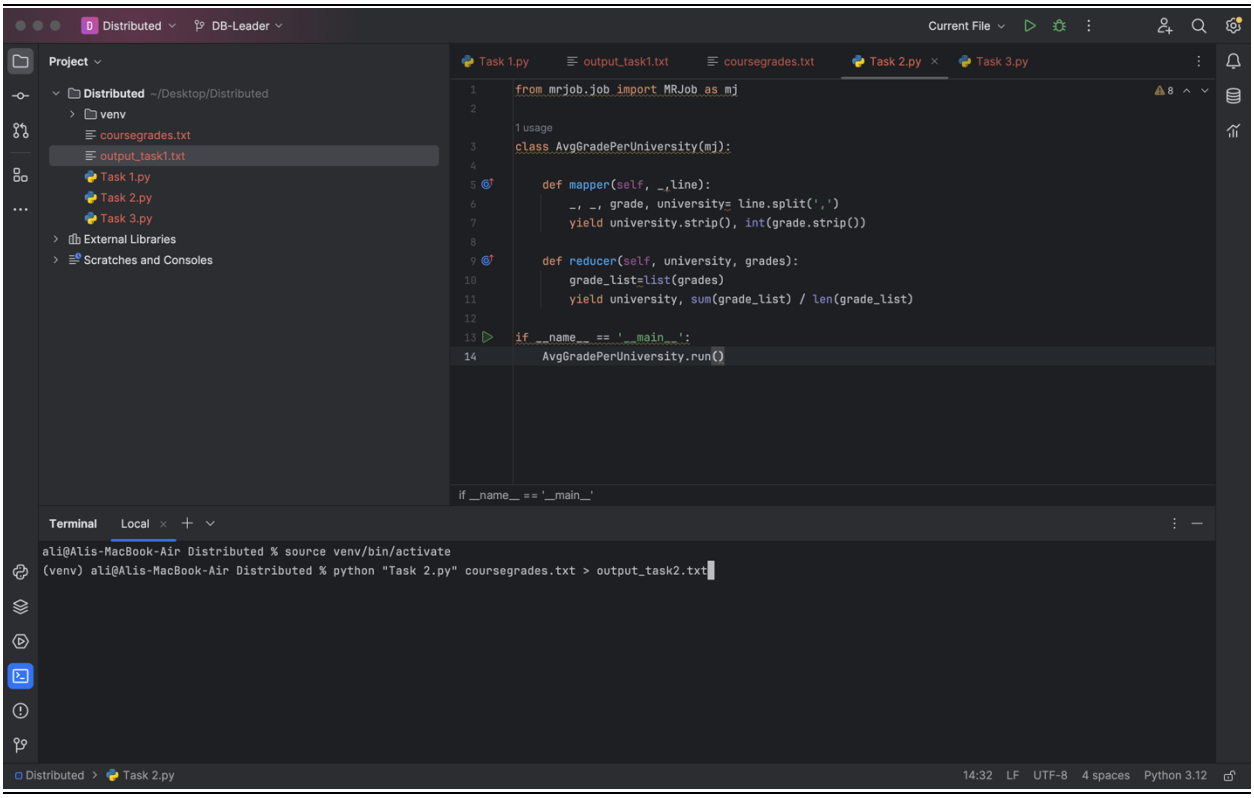

Distributed
DB-Leader
Current File
Project
Distributed ~/Desktop/Distributed
venv
coursegrades.txt
Task 1.py
Task 2.py
Task 3.py
External Libraries
Scratches and Consoles
Task 1.py
coursegrades.txt
Task 2.py
Task 3.py
from mrjob.job import MRJob as mj
1 usage
class AvgGradePerCourse(mj):
def mapper(self, _, line):
_, course, grade, _= line.split(',')
yield course.strip(), int(grade.strip())
def reducer(self, course, grades):
grade_list=list(grades)
yield course, sum(grade_list)/len(grade_list)
if __name__ == '__main__':
AvgGradePerCourse.run()
if __name__ == '__main__'
Distributed
Task 1.py
14:28
LF
UTF-8
4 spaces
Python 3.12

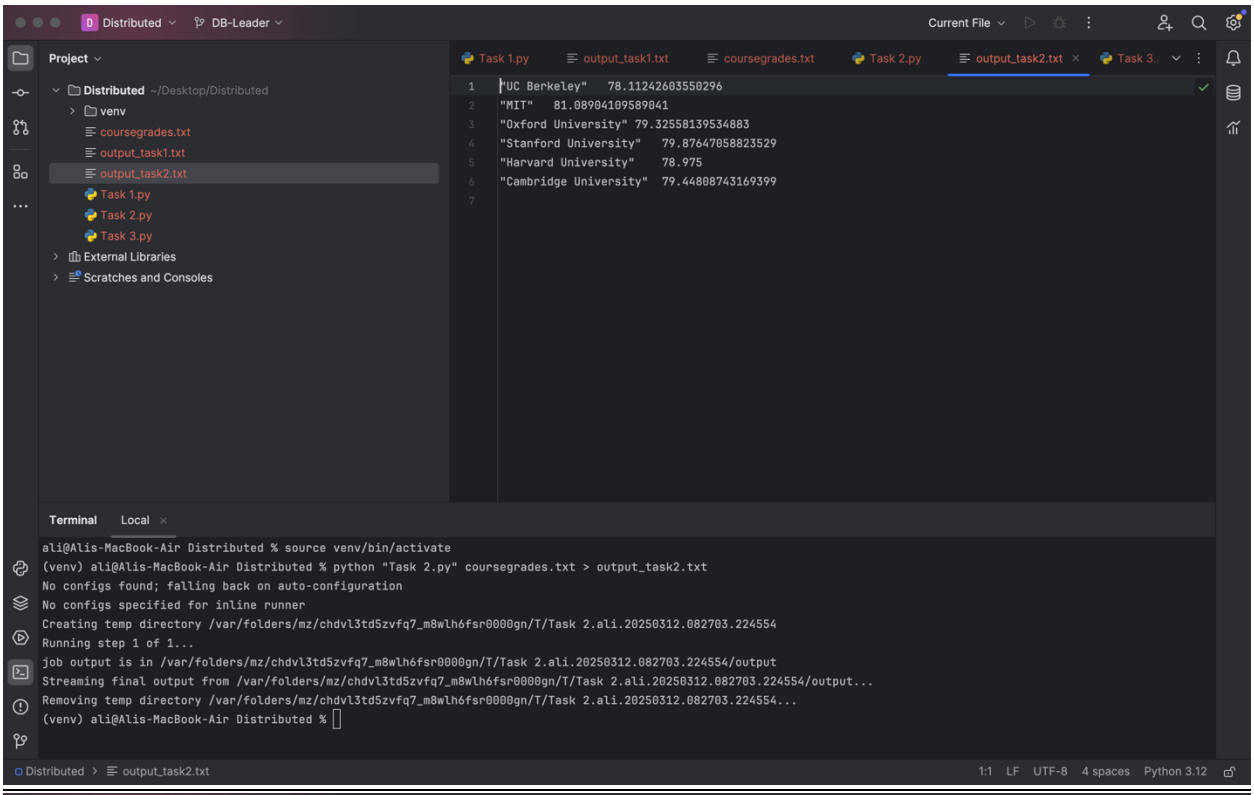

Distributed
DB-Leader
Current File
Project
Distributed ~/Desktop/Distributed
venv
coursegrades.txt
Task 1.py
Task 2.py
Task 3.py
External Libraries
Scratches and Consoles
Task 1.py
coursegrades.txt
Task 2.py
Task 3.py
from mrjob.job import MRJob as mj
1 usage
class AvgGradePerCourse(mj):
def mapper(self, _, line):
_, course, grade, _= line.split(',')
yield course.strip(), int(grade.strip())
def reducer(self, course, grades):
grade_list=list(grades)
yield course, sum(grade_list)/len(grade_list)
if __name__ == '__main__':
AvgGradePerCourse.run()
if __name__ == '__main__'
Terminal
Local
ali@Alis-MacBook-Air Distributed % source venv/bin/activate
(venv) ali@Alis-MacBook-Air Distributed %
Distributed
Task 1.py
14:28
LF
UTF-8
4 spaces
Python 3.12

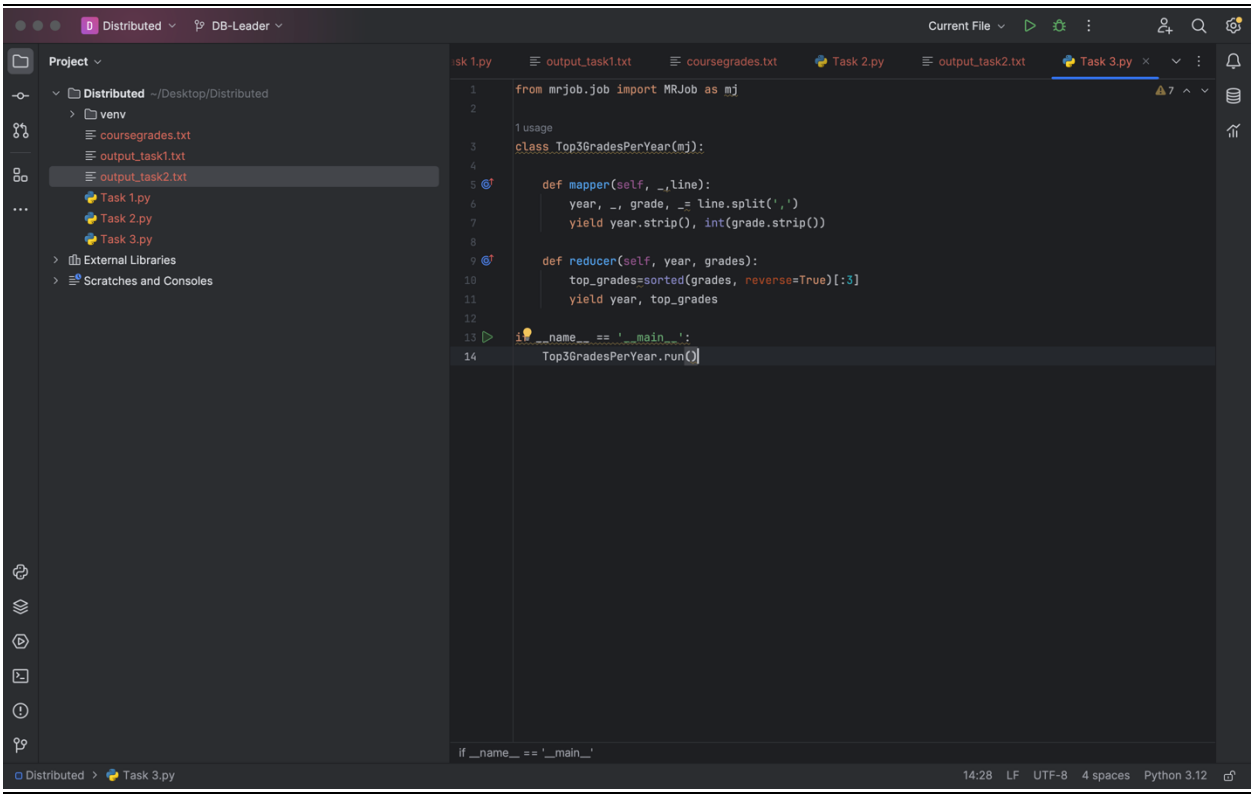

Distributed
DB-Leader
Current File
Project
Distributed ~/Desktop/Distributed
venv
coursegrades.txt
Task 1.py
Task 2.py
Task 3.py
External Libraries
Scratches and Consoles
Task 1.py
coursegrades.txt
Task 2.py
Task 3.py
from mrjob.job import MRJob as mj
1 usage
class AvgGradePerCourse(mj):
def mapper(self, _, line):
_, course, grade, _= line.split(',')
yield course.strip(), int(grade.strip())
def reducer(self, course, grades):
grade_list=list(grades)
yield course, sum(grade_list)/len(grade_list)
if __name__ == '__main__':
AvgGradePerCourse.run()
if __name__ == '__main__'
Terminal
Local
ali@Alis-MacBook-Air Distributed % source venv/bin/activate
(venv) ali@Alis-MacBook-Air Distributed % python "Task 1.py" coursegrades.txt > output_task1.txt
Distributed
Task 1.py
14:28
LF
UTF-8
4 spaces
Python 3.12

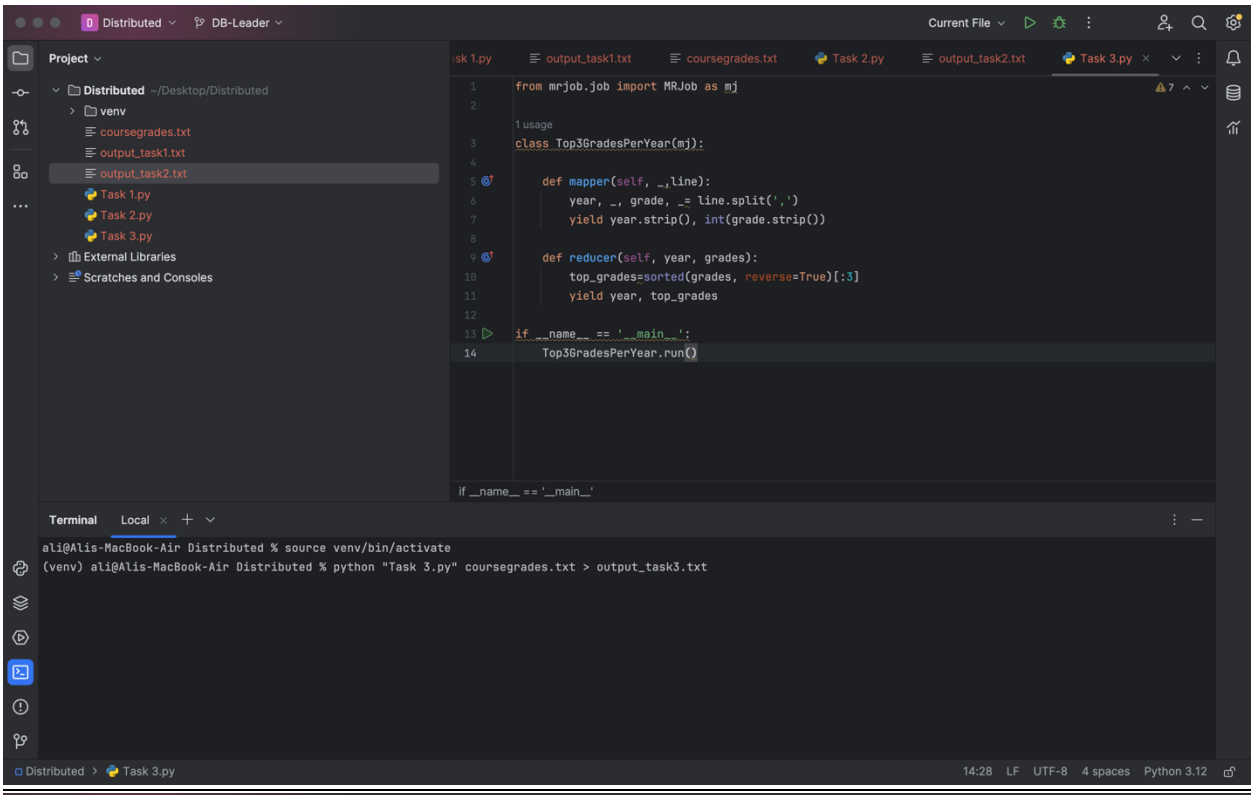

Distributed
DB-Leader
Current File
Project
Distributed ~/Desktop/Distributed
venv
coursegrades.txt
output_task1.txt
Task 1.py
Task 2.py
Task 3.py
External Libraries
Scratches and Consoles
Task 1.py
output_task1.txt
coursegrades.txt
Task 2.py
Task 3.py
"Software Engineering" 78.08695652173913
"Cyber Security" 78.58333333333333
"Data Structures" 81.45341614906832
"Machine Learning" 79.43125
"Computer Vision" 79.73958333333333
"Artificial Intelligence" 79.31012658227849
Terminal
Local
ali@Alis-MacBook-Air Distributed % source venv/bin/activate
(venv) ali@Alis-MacBook-Air Distributed % python "Task 1.py" coursegrades.txt > output_task1.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 1.ali.20250312.082606.621532
Running step 1 of 1...
job output is in /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 1.ali.20250312.082606.621532/output
Streaming final output from /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 1.ali.20250312.082606.621532/output...
Removing temp directory /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 1.ali.20250312.082606.621532...
(venv) ali@Alis-MacBook-Air Distributed %
Distributed
output_task1.txt
1:1
LF
UTF-8
4 spaces
Python 3.12

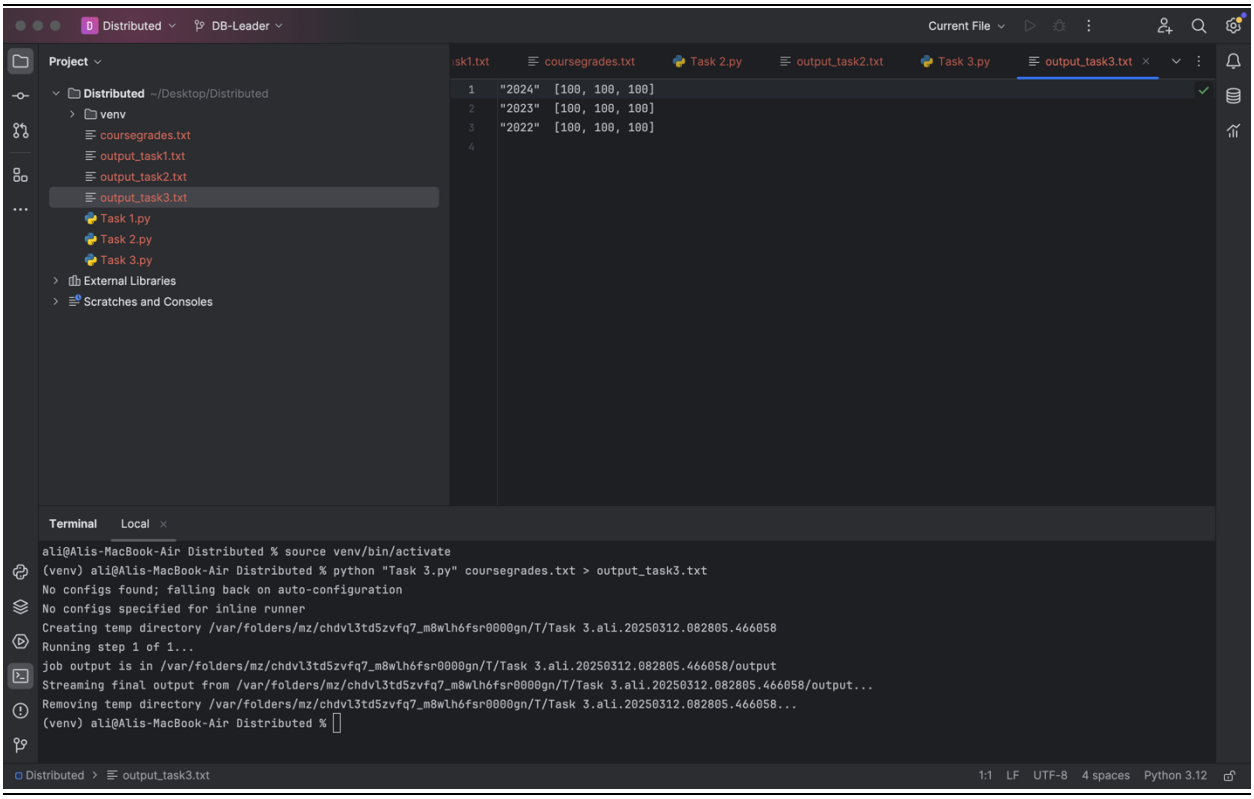

Distributed
DB-Leader
Current File
Project
Distributed ~/Desktop/Distributed
venv
coursegrades.txt
output_task1.txt
Task 1.py
Task 2.py
Task 3.py
External Libraries
Scratches and Consoles
Task 1.py
output_task1.txt
coursegrades.txt
Task 2.py
Task 3.py
from mrjob.job import MRJob as mj
1 usage
class AvgGradePerUniversity(mj):
def mapper(self, _,line):
_, _, grade, university= line.split(',')
yield university.strip(), int(grade.strip())
def reducer(self, university, grades):
grade_list=list(grades)
yield university, sum(grade_list) / len(grade_list)
if __name__ == '__main__':
AvgGradePerUniversity.run()
if __name__ == '__main__'
Distributed
Task 2.py
14:32
LF
UTF-8
4 spaces
Python 3.12


Distributed
DB-Leader
Current File
Project
Distributed ~/Desktop/Distributed
venv
coursegrades.txt
output_task1.txt
Task 1.py
Task 2.py
Task 3.py
External Libraries
Scratches and Consoles
Task 1.py
output_task1.txt
coursegrades.txt
Task 2.py
Task 3.py
from mrjob.job import MRJob as mj
1 usage
class AvgGradePerUniversity(mj):
def mapper(self, _,line):
_, _, grade, university= line.split(',')
yield university.strip(), int(grade.strip())
def reducer(self, university, grades):
grade_list=list(grades)
yield university, sum(grade_list) / len(grade_list)
if __name__ == '__main__':
AvgGradePerUniversity.run()
if __name__ == '__main__'
Terminal
Local
ali@Alis-MacBook-Air Distributed % source venv/bin/activate
(venv) ali@Alis-MacBook-Air Distributed % python "Task 2.py" coursegrades.txt > output_task2.txt
Distributed
Task 2.py
14:32
LF
UTF-8
4 spaces
Python 3.12


Distributed
DB-Leader
Current File
Project
Distributed ~/Desktop/Distributed
venv
coursegrades.txt
output_task1.txt
output_task2.txt
Task 1.py
Task 2.py
Task 3.py
External Libraries
Scratches and Consoles
Task 1.py
output_task1.txt
coursegrades.txt
Task 2.py
output_task2.txt
Task 3
"UC Berkeley" 78.11242603550296
"MIT" 81.08904109589041
"Oxford University" 79.32558139534883
"Stanford University" 79.87647058823529
"Harvard University" 78.975
"Cambridge University" 79.44808743169399
Terminal
Local
ali@Alis-MacBook-Air Distributed % source venv/bin/activate
(venv) ali@Alis-MacBook-Air Distributed % python "Task 2.py" coursegrades.txt > output_task2.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 2.ali.20250312.082703.224554
Running step 1 of 1...
job output is in /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 2.ali.20250312.082703.224554/output
Streaming final output from /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 2.ali.20250312.082703.224554/output...
Removing temp directory /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 2.ali.20250312.082703.224554...
(venv) ali@Alis-MacBook-Air Distributed %
Distributed
output_task2.txt
1:1
LF
UTF-8
4 spaces
Python 3.12


Distributed
DB-Leader
Current File
Project
Distributed ~/Desktop/Distributed
venv
coursegrades.txt
output_task1.txt
output_task2.txt
Task 1.py
Task 2.py
Task 3.py
External Libraries
Scratches and Consoles
output_task1.txt
coursegrades.txt
Task 2.py
output_task2.txt
Task 3.py
from mrjob.job import MRJob as mj
1 usage
class Top3GradesPerYear(mj):
def mapper(self, _,line):
year, _, grade, _= line.split(',')
yield year.strip(), int(grade.strip())
def reducer(self, year, grades):
top_grades=sorted(grades, reverse=True)[:3]
yield year, top_grades
if __name__ == '__main__':
Top3GradesPerYear.run()
if __name__ == '__main__'
Distributed
Task 3.py
14:28
LF
UTF-8
4 spaces
Python 3.12


Distributed
DB-Leader
Current File
Project
Distributed ~/Desktop/Distributed
venv
coursegrades.txt
output_task1.txt
output_task2.txt
Task 1.py
Task 2.py
Task 3.py
External Libraries
Scratches and Consoles
Task 3.py
output_task1.txt
coursegrades.txt
Task 2.py
output_task2.txt
from mrjob.job import MRJob as mj
1 usage
class Top3GradesPerYear(mj):
def mapper(self, _,line):
year, _, grade, _= line.split(',')
yield year.strip(), int(grade.strip())
def reducer(self, year, grades):
top_grades=sorted(grades, reverse=True)[:3]
yield year, top_grades
if __name__ == '__main__':
Top3GradesPerYear.run()
if __name__ == '__main__'
Terminal
Local
ali@Alis-MacBook-Air Distributed % source venv/bin/activate
(venv) ali@Alis-MacBook-Air Distributed % python "Task 3.py" coursegrades.txt > output_task3.txt
Distributed
Task 3.py
14:28
LF
UTF-8
4 spaces
Python 3.12


Distributed
DB-Leader
Current File
Project
Distributed ~/Desktop/Distributed
venv
coursegrades.txt
output_task1.txt
output_task2.txt
output_task3.txt
Task 1.py
Task 2.py
Task 3.py
External Libraries
Scratches and Consoles
coursegrades.txt
Task 2.py
output_task2.txt
Task 3.py
output_task3.txt
"2024" [100, 100, 100]
"2023" [100, 100, 100]
"2022" [100, 100, 100]
Terminal
Local
ali@Alis-MacBook-Air Distributed % source venv/bin/activate
(venv) ali@Alis-MacBook-Air Distributed % python "Task 3.py" coursegrades.txt > output_task3.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 3.ali.20250312.082805.466058
Running step 1 of 1...
job output is in /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 3.ali.20250312.082805.466058/output
Streaming final output from /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 3.ali.20250312.082805.466058/output...
Removing temp directory /var/folders/mz/chdvl3td5zvfq7_m8wlh6fsr0000gn/T/Task 3.ali.20250312.082805.466058...
(venv) ali@Alis-MacBook-Air Distributed %
Distributed
output_task3.txt
1:1
LF
UTF-8
4 spaces
Python 3.12