MapReduce Programs

Maimoona Khilji

Institute of Management Science

Course Code: Big Data Programming

Imran Ahmad Mughal

11th November, 2021

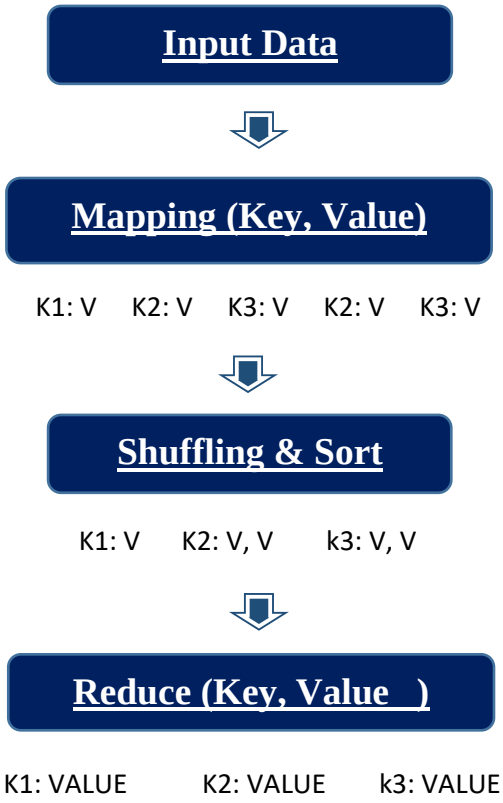
MapReduce

MapReduce is used for writing applications that can process huge amount of data on large clusters. MapReduce basically consists of two functions:

- **Mapper:**
  - Mapper function is used to coverts the input data into pairs(key, value)
- **Reducer:**
  - Reducer function is used to process the key's value and outputs the desired result.

Steps:

1. Input the source data.
2. Mapper **( )** function maps the input data in form of key-value pairs.
3. Shuffle and sort the data generated by mapper, to make the groups of equivalent keys easy.
4. Reducer ( ) function aggregates the value according to keys, to yield key-value pair in output.



1. Property Dataset

Problem: How many times has a particular type of property been sold?

Solution:

i. Input the **property** data set.

ii. Through **mapper** ( ) function, select the column **"Type"** as Key and assign **1** as value to each key as

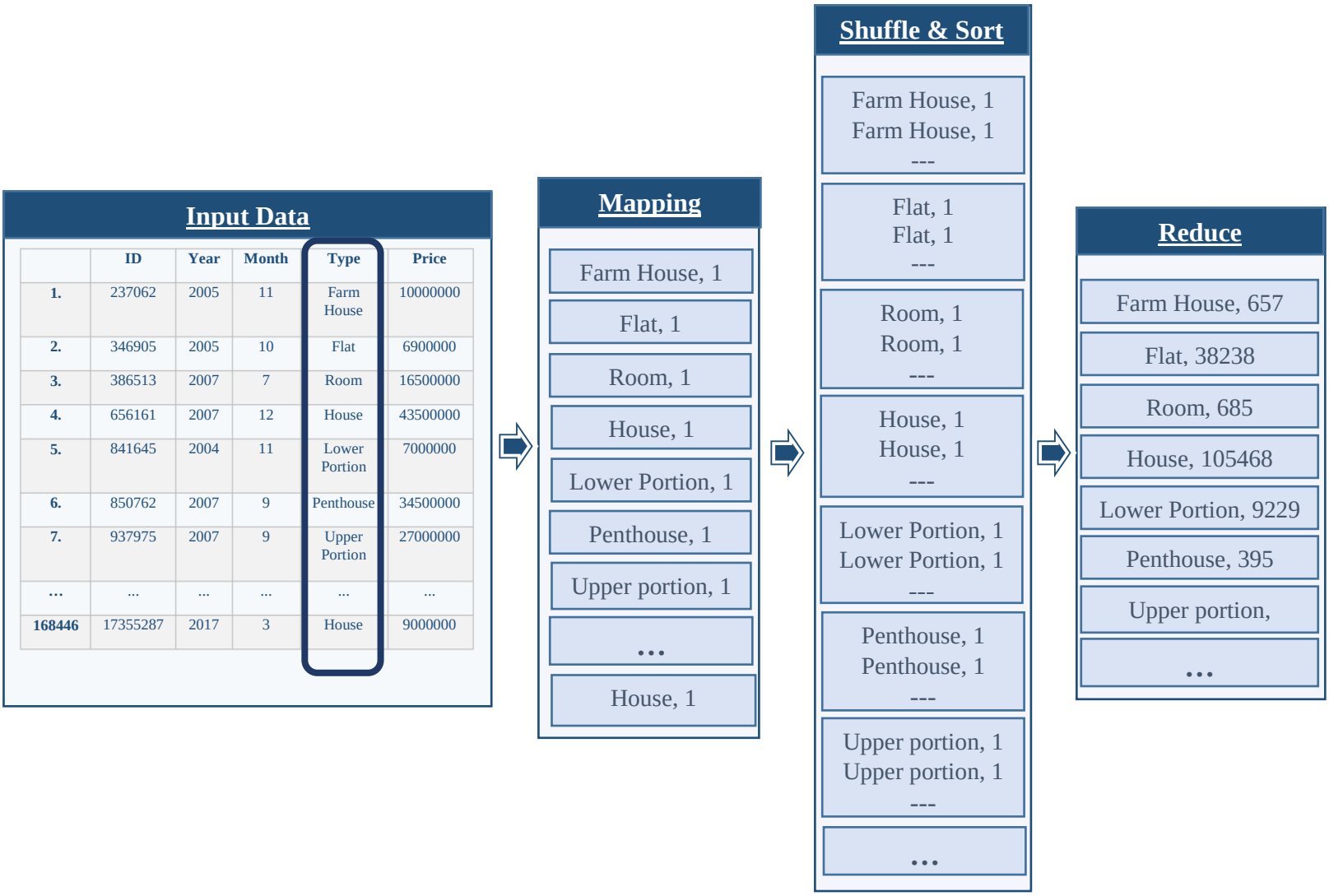
(Key: value = Type: 1).

iii. **Shuffle and sort** the key-value pair and then group them on the basis of key.

iv. Through **Reducer** ( ) function, calculate the sum of occurrence of each type (by addition of value=1 assigned to each key) and then yield the result in key-value pair where:

a. **Key** represents the **type of property**

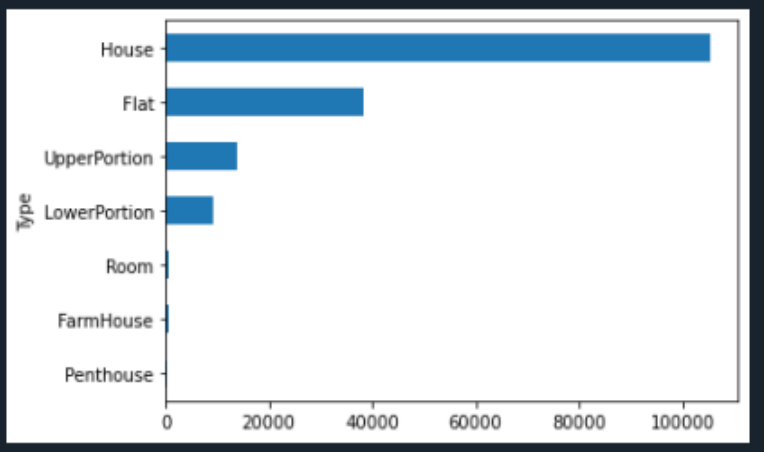
b. **Value** represents count of each type.



Output (Spyder IDE)

property.py

```
from mrjob.job import MRJob

#mapreduce

class PropertyType(MRJob):

    def mapper(self, key, line):
        ID,Year,Month,Type,Price = line.split(",")
        yield Type, 1

    def reducer(self,Type,occurence):
        yield Type,sum(occurence)

if __name__ == '__main__':
    PropertyType.run()


#plot
import pandas as pd

df= pd.read_csv("Property.csv")
df.columns=['ID','Year','Month','Type','Price']

x = df[df.columns[3]]

x = df.groupby('Type')['ID'].count().sort_values()

x.plot(kind='barh',ylabel="Type")
```



Console 3/A

```
In [1]: !python property.py Property.csv
"FarmHouse" 657
"Flat"  38238
"House" 105468
"LowerPortion"  9229
"Penthouse" 395
"Room"  685
"UpperPortion"  13774
```

Result

A number of particular type of property been sold is:

a) **Farmhouse** has been sold **657** times
b) **Flat** has been sold **38238** times
c) **House** has been sold **105468** times
d) **Lower Portion** has been sold **9229** times
e) **Penthouse** has been sold **395** times
f) **Room** has been sold **685** times
g) **Upper Portion** has been sold **13774** times

2. Property Dataset

Problem: In what year were most of the properties sold?

Solution:

i. Input the **property** data set.

ii. Through **mapper** ( ) function, select the column **"Year"** as Key and assign **1** as value to each key as

(Key: value = Type: 1).

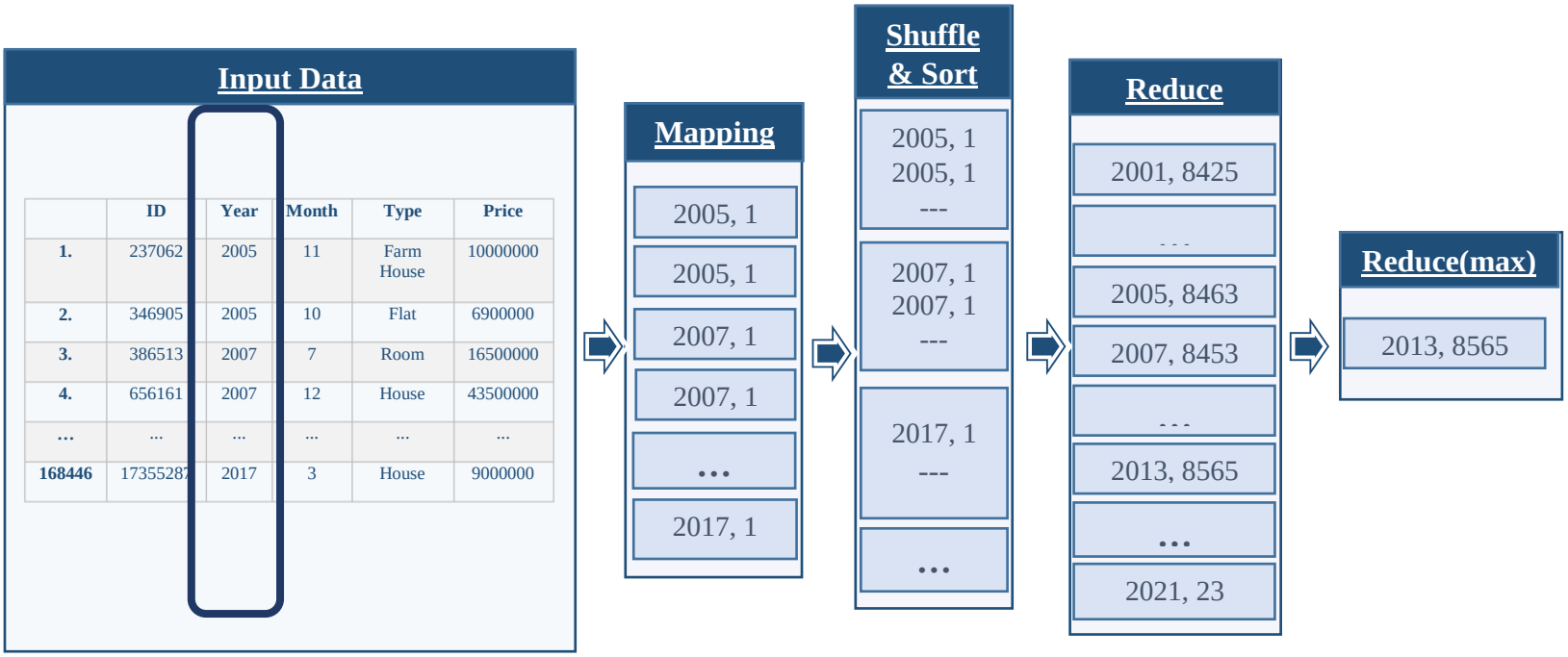
iii. **Shuffle and sort** the key-value pair and then group them on the basis of key.

iv. Through **Reducer** ( ) function, calculate the sum of occurrence of each type (by addition of value=1 assigned to each key) and then yield the result in key-value pair where:

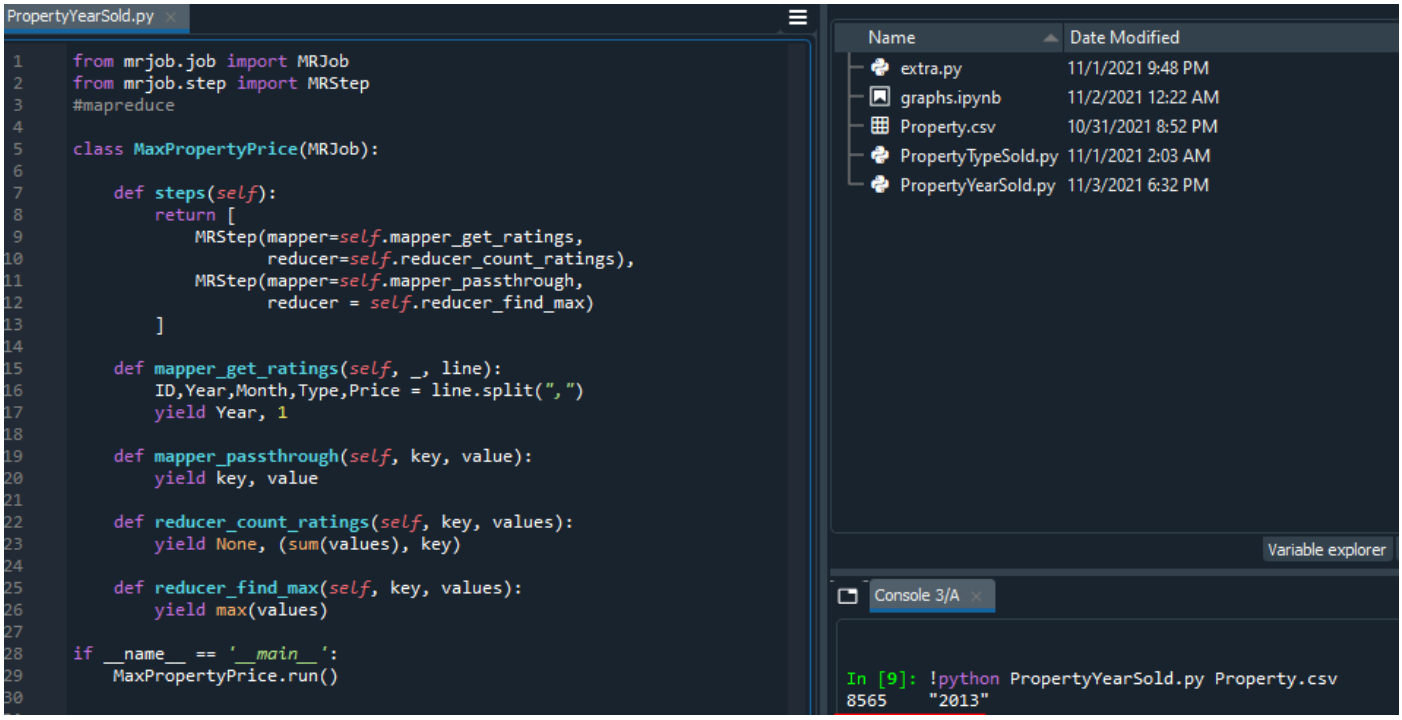
a. **Key** represents the **type of property**

b. **Value** represents count of each type.

v. These values is then reduced by finding the **maximum count** and then yield the year with maximum property sold.



Output (Spyder IDE)



Result

Most of the properties were sold in **2013.**

As 8565 Properties were sold in 2013 which is the maximum count.)

3. Shoe Shop Dataset

Problem: Earning (Total Sale price) per year?

Solution:

i. Input the **Shoe Shop** data set.

ii. Through **mapper** ( ) function, select the column **"Year"** as Key and **"Sale Price"** as value to each key as

(Key: value = Year: Sale_Price).

iii. **Shuffle and sort** the key-value pair and then group them on the basis of key.

iv. Through **Reducer** ( ) function, calculate the sum of Sale price for each year and then yield the result in key-value pair where:

a. **Key** represents the **year**

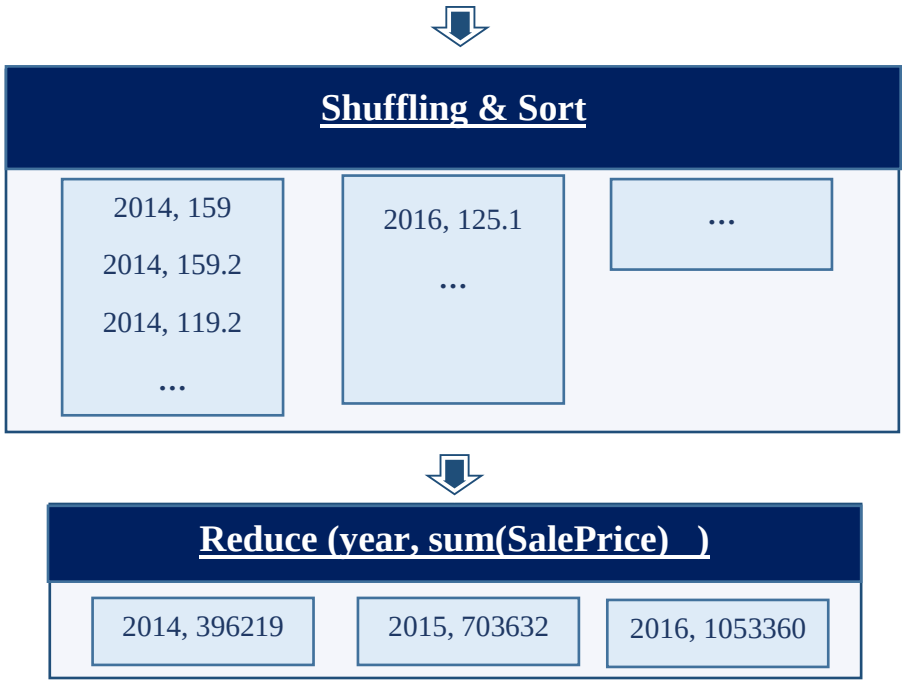
b. Value represents **Total Sale price.**

Input Data

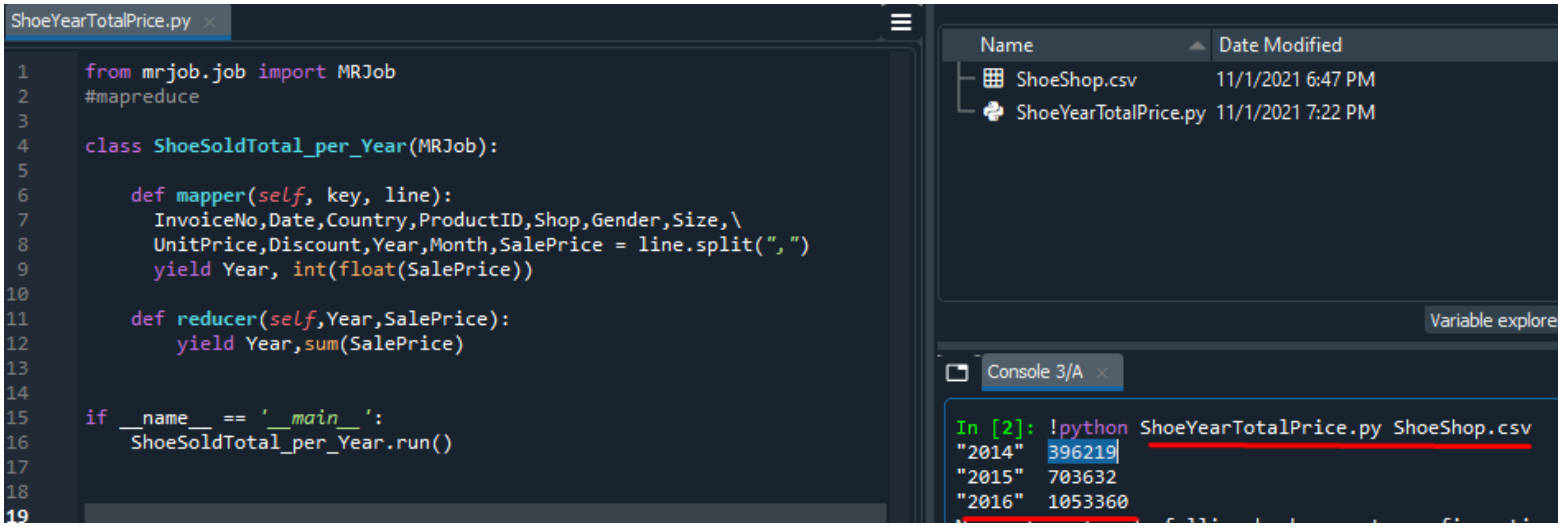
	InvoiceNo	Date	Country	ProductID	Shop	Gender	Size	UnitPrice	Discount	Year	Month	SalePrice
1.	52389	1/1/2014	United Kingdom	2152	UK2	Male	11	159	0%	2014	1	159
2.	52390	1/1/2014	United States	2230	US15	Male	11.5	199	20%	2014	1	159.2
3.	52391	1/1/2014	Canada	2160	CAN7	Male	9.5	149	20%	2014	1	119.2
...	---	---	---	---	---	---	---	---	---	---	---	---
14967	65777	12/31/2016	Germany	2156	GER1	Female	6.5	139	10%	2016	12	125.1

Mapping (Key=Year, Value= Sale Price)

2014, 159	2014, 159.2	2014, 119.2	2016, 125.1	...



Output (Spyder IDE)



Result

Total Sale price per year is:

a) Total Sale price was **3,96,219** in **2014**
b) Total Sale price was **7,03,632** in **2015**
c) Total Sale price was **1,053,360** in **2016**

4. Shoe Shop Dataset

Problem: How many shoes have been sold in each country?

Solution:

i. Input the **Shoe Shop** data set.

ii. Through **mapper** ( ) function, select the column **"Country"** as Key and assign **1** as value to each key as

(Key: value = Country: 1).

iii. **Shuffle and sort** the key-value pair and then group them on the basis of key.

iv. Through **Reducer** ( ) function, calculate the sum of occurrence of each country (by addition of value=1 assigned to each key) and then yield the result in key-value pair where:

a. **Key** represents the **Country**

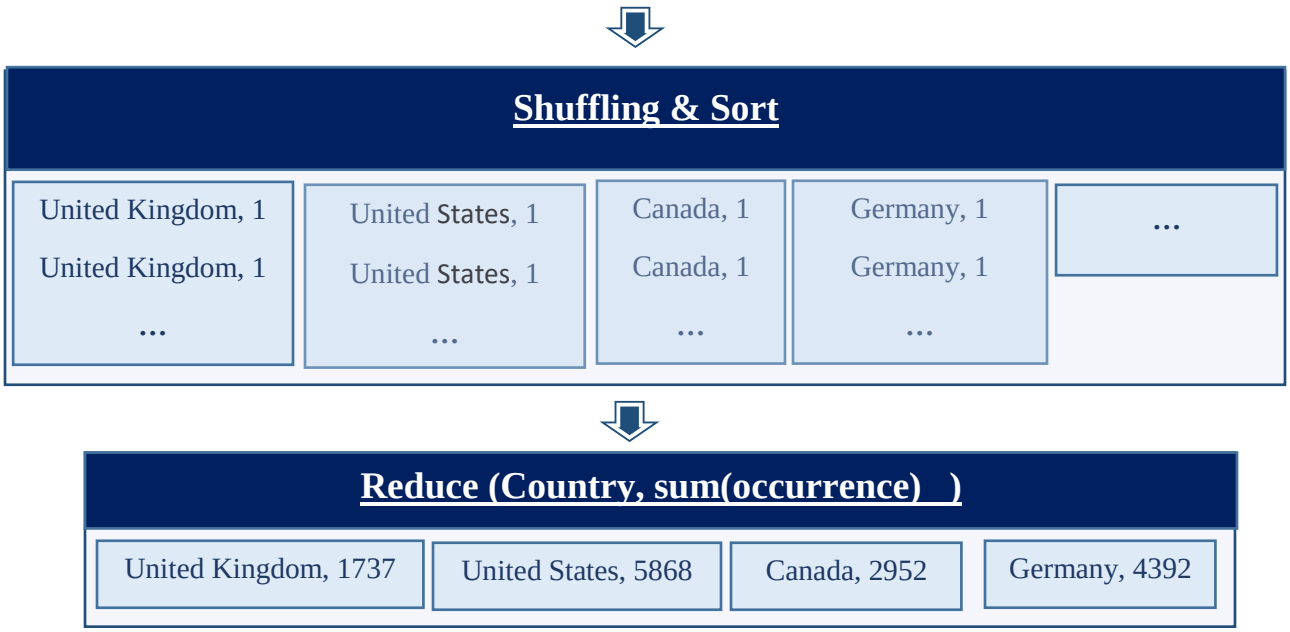
b. **Value** represents count of each Country.

Input Data

	InvoiceNo	Date	Country	ProductID	Shop	Gender	Size	UnitPrice	Discount	Year	Month	SalePrice
1.	52389	1/1/2014	United Kingdom	2152	UK2	Male	11	159	0%	2014	1	159
2.	52390	1/1/2014	United States	2230	US15	Male	11.5	199	20%	2014	1	159.2
3.	52391	1/1/2014	Canada	2160	CAN7	Male	9.5	149	20%	2014	1	119.2
...	---	---	---	---	---	---	---	---	---	---	---	---
14967	65777	12/31/2016	Germany	2156	GER1	Female	6.5	139	10%	2016	12	125.1

Mapping (Key=Country, Value= 1)

United Kingdom, 1	United States, 1	Canada, 1	...	Germany, 1



Output (Spyder IDE)

Result

The number of shoes have been sold in each country are:

a) In **Canada,** 2,952 shoes have been sold.
b) In **Germany,** 4,392 shoes have been sold.
c) In **United Kingdom,** 1,737 shoes have been sold.
d) In **United States,** 5,886 shoes have been sold.

5. Shoe Shop Dataset

Problem: How much has each country earned by selling shoes?

Solution:

i. Input the **Shoe Shop** data set.

ii. Through **mapper** ( ) function, select the column **"Country"** as Key and **"Sale Price"** as value to each key as

(Key: value = Country: Sale_Price).

iii. **Shuffle and sort** the key-value pair and then group them on the basis of key.

iv. Through **Reducer** ( ) function, calculate the sum of Sale price for each Country and then yield the result in key-value pair where:

a. **Key** represents the **Country**

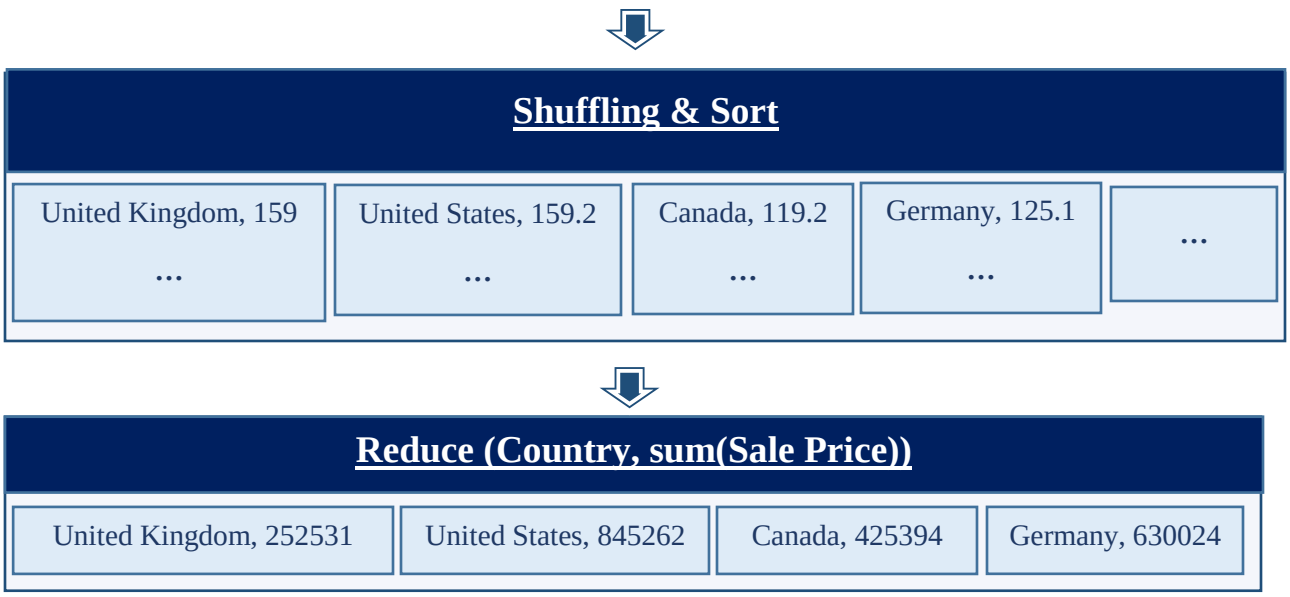
b. Value represents **Total Sale price.**

Input Data

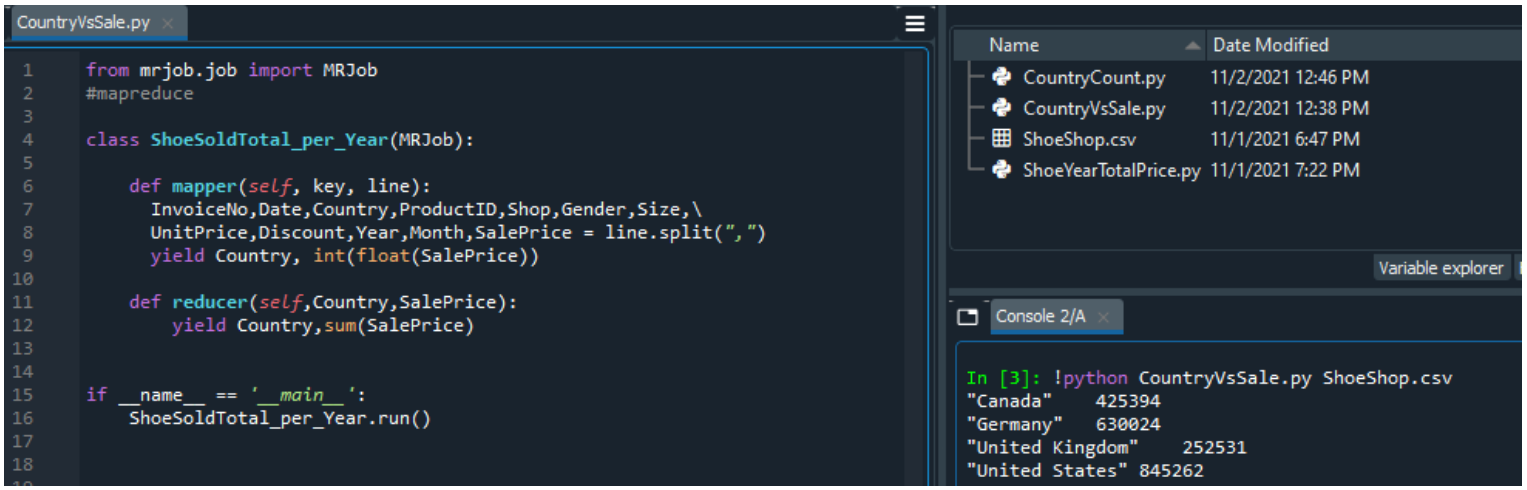
	InvoiceNo	Date	Country	ProductID	Shop	Gender	Size	UnitPrice	Discount	Year	Month	SalePrice
1.	52389	1/1/2014	United Kingdom	2152	UK2	Male	11	159	0%	2014	1	159
2.	52390	1/1/2014	United States	2230	US15	Male	11.5	199	20%	2014	1	159.2
3.	52391	1/1/2014	Canada	2160	CAN7	Male	9.5	149	20%	2014	1	119.2
...	---	---	---	---	---	---	---	---	---	---	---	---
14967	65777	12/31/2016	Germany	2156	GER1	Female	6.5	139	10%	2016	12	125.1

Mapping (Key=Country, Value= Sale Price)

United Kingdom, 159	United States, 159.2	Canada, 119.2	Germany, 125.1	…



Output (Spyder IDE)



Result

The earning of each country is given below:

a) **Canada** has earned 4,25,394
b) **Germany,** has earned 6,30,024
c) **United Kingdom,** has earned 2,52,531
d) **United States,** has earned 8,45,262

6. Shoe Shop Dataset

Problem: How many shops are there for Males or Females?

Solution:

i. Input the **Shoe Shop** data set.

ii. Through **mapper** ( ) function, select the column **"Gender"** as Key and assign **1** as value to each key as

(Key: value = Gender: 1).

iii. **Shuffle and sort** the key-value pair and then group them on the basis of key.

iv. Through **Reducer** ( ) function, calculate the sum of occurrence of each gender (by addition of value=1 assigned to each key) and then yield the result in key-value pair where:

a. **Key** represents the **Gender**

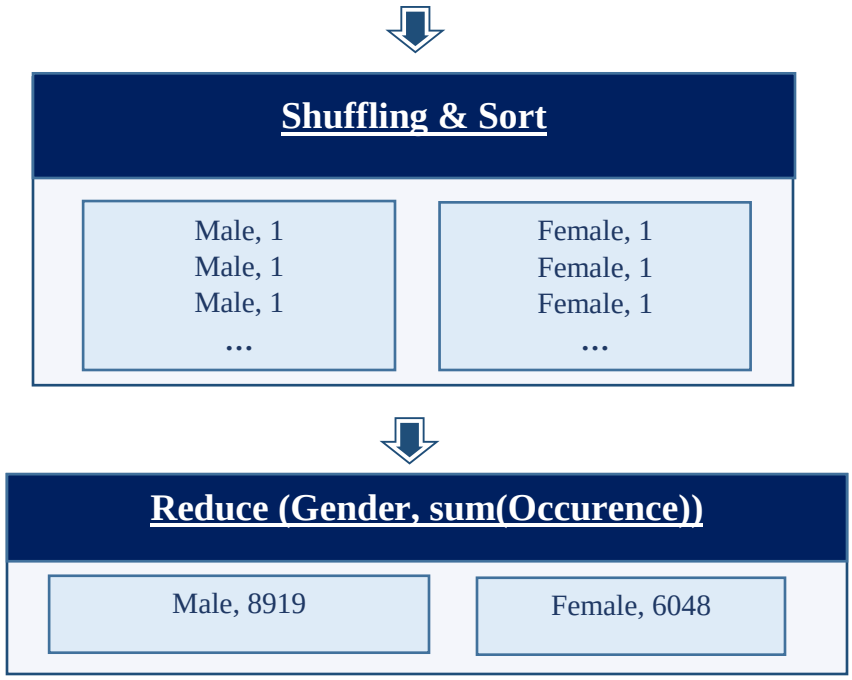
b. **Value** represents count of each Gender.

Input Data

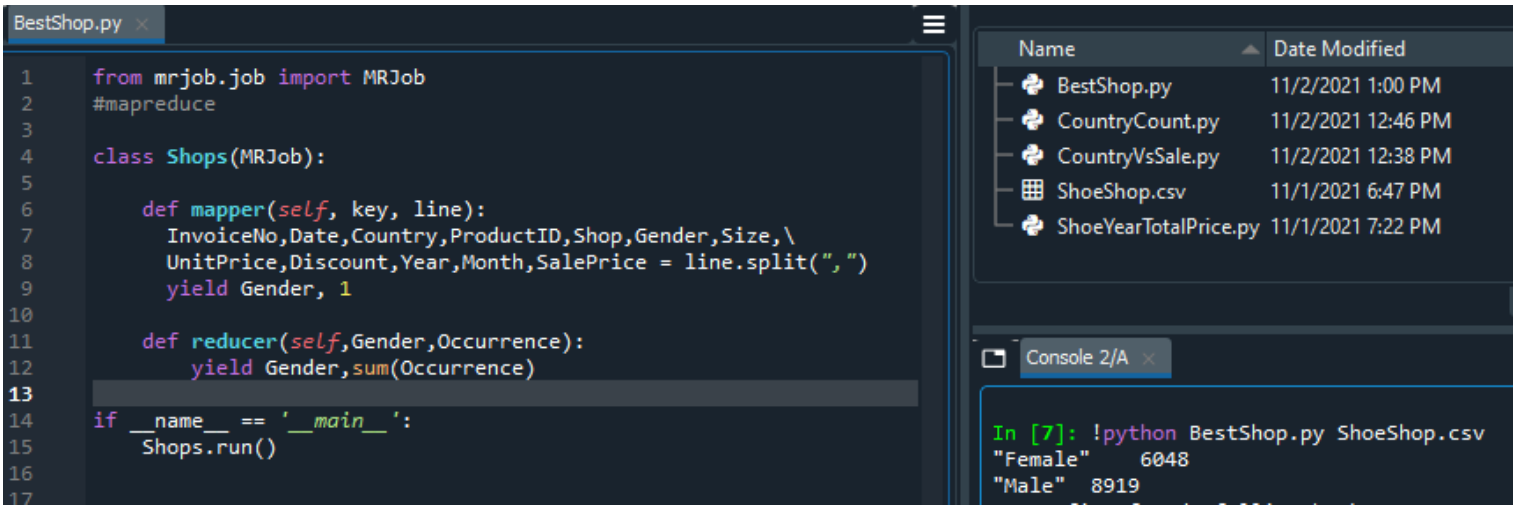
	InvoiceNo	Date	Country	ProductID	Shop	Gender	Size	UnitPrice	Discount	Year	Month	SalePrice
1.	52389	1/1/2014	United Kingdom	2152	UK2	Male	11	159	0%	2014	1	159
2.	52390	1/1/2014	United States	2230	US15	Male	11.5	199	20%	2014	1	159.2
3.	52391	1/1/2014	Canada	2160	CAN7	Male	9.5	149	20%	2014	1	119.2
...	---	---	---	---	---	---	---	---	---	---	---	---
14967	65777	12/31/2016	Germany	2156	GER1	Female	6.5	139	10%	2016	12	125.1

Mapping (Key=Gender, Value= 1)

Male, 1	Male, 1	Male, 1	…	Female, 1



Output (Spyder IDE)



Result

a) There are **6,048** shops for Female
b) There are **8,919** shops for Male

7. Shoe Shop Dataset

Problem: How many shoes have been sold per year?

Solution:

i. Input the **Shoe Shop** data set.

ii. Through **mapper** ( ) function, select the column **"Year"** as Key and assign **1** as value to each key as

(Key: value = Year: 1).

iii. **Shuffle and sort** the key-value pair and then group them on the basis of key.

iv. Through **Reducer** ( ) function, calculate the sum of occurrence of each Year (by addition of value=1 assigned to each key) and then yield the result in key-value pair where:

a. **Key** represents the **Year**

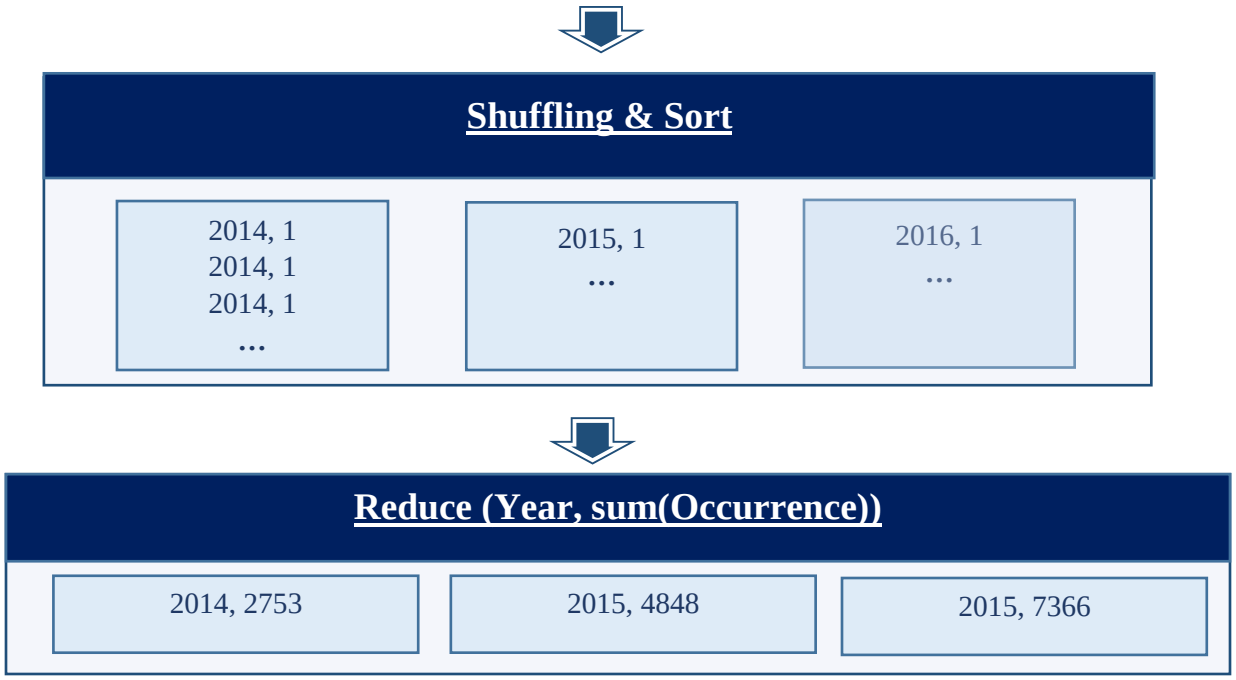
b. **Value** represents count of each Year.

Input Data

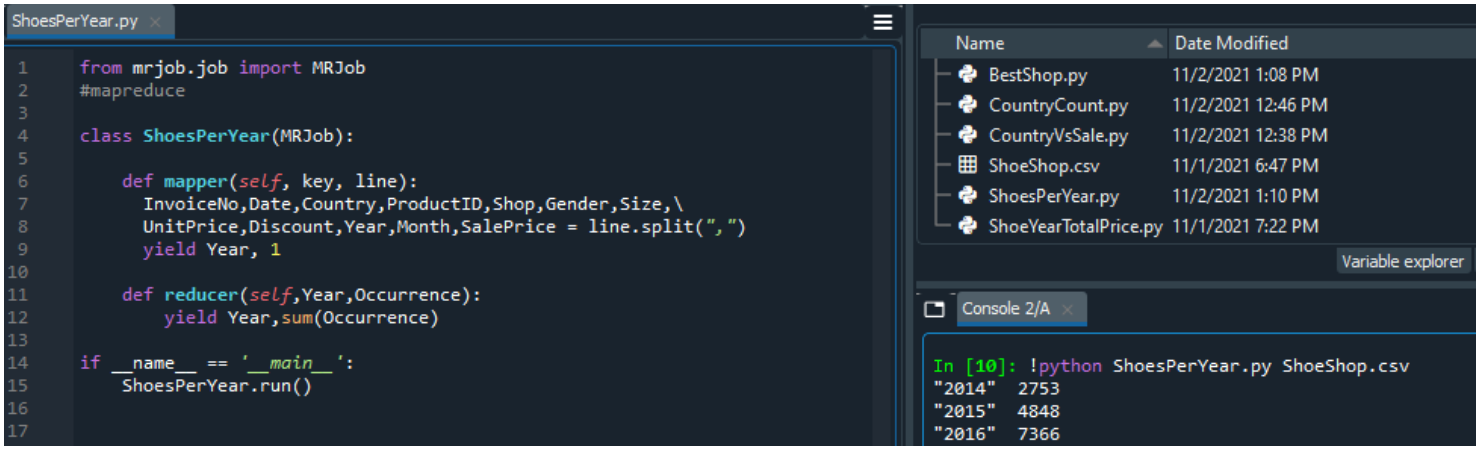
	InvoiceNo	Date	Country	ProductID	Shop	Gender	Size	UnitPrice	Discount	Year	Month	SalePrice
1.	52389	1/1/2014	United Kingdom	2152	UK2	Male	11	159	0%	2014	1	159
2.	52390	1/1/2014	United States	2230	US15	Male	11.5	199	20%	2014	1	159.2
3.	52391	1/1/2014	Canada	2160	CAN7	Male	9.5	149	20%	2014	1	119.2
...	---	---	---	---	---	---	---	---	---	---	---	---
14967	65777	12/31/2016	Germany	2156	GER1	Female	6.5	139	10%	2016	12	125.1

Mapping (Key=Year, Value= 1)

2014, 1	2014, 1	2014, 1	...	2016, 1



Output (Spyder IDE)



Result

The number of shoes have been sold per year is given below:

a) In **2014**, the number of shoes have been sold is **2,753.**
b) In **2015**, the number of shoes have been sold is **4,848**.
c) In **2016**, the number of shoes have been sold is **7,366**.

8. Shoe Shop Dataset

Problem: Average discount offered by each country?

Solution:

i. Input the **Shoe Shop** data set.

ii. Through **mapper** ( ) function, select the column **"Country"** as Key and **"Discount"** as value to each key as

(Key: value = Country: Discount).

iii. **Shuffle and sort** the key-value pair and then group them on the basis of key.

iv. Through **Reducer** ( ) function, calculate the average of Discount for each Country and then yield the result in key-value pair where:

a. **Key** represents the **Country**

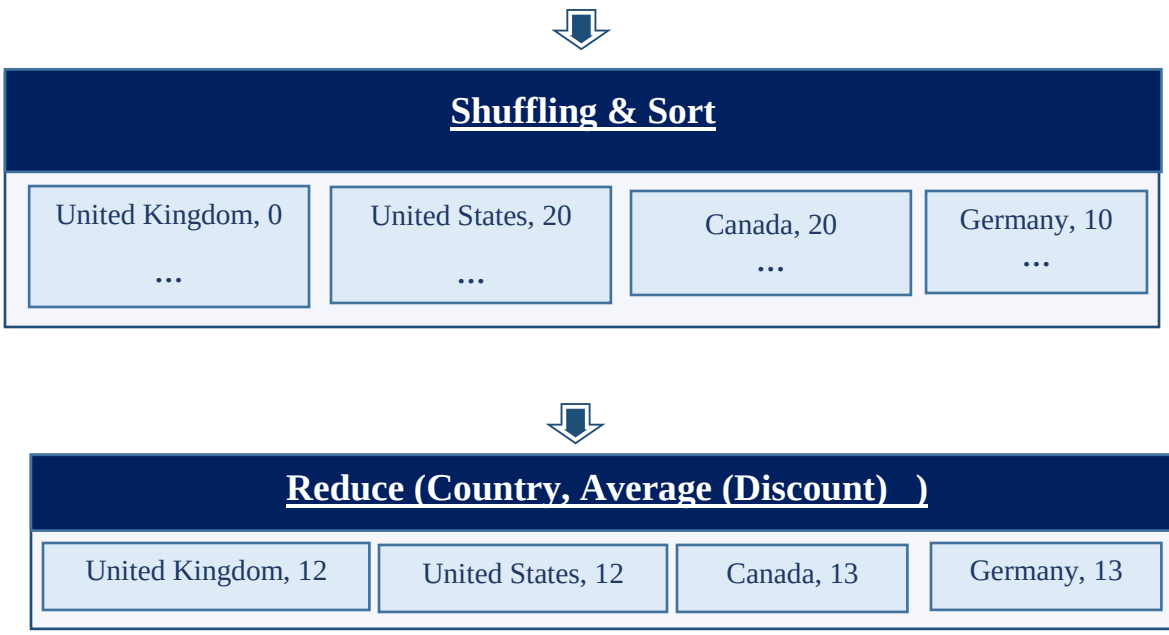
b. Value represents **Average Discount.**

Input Data

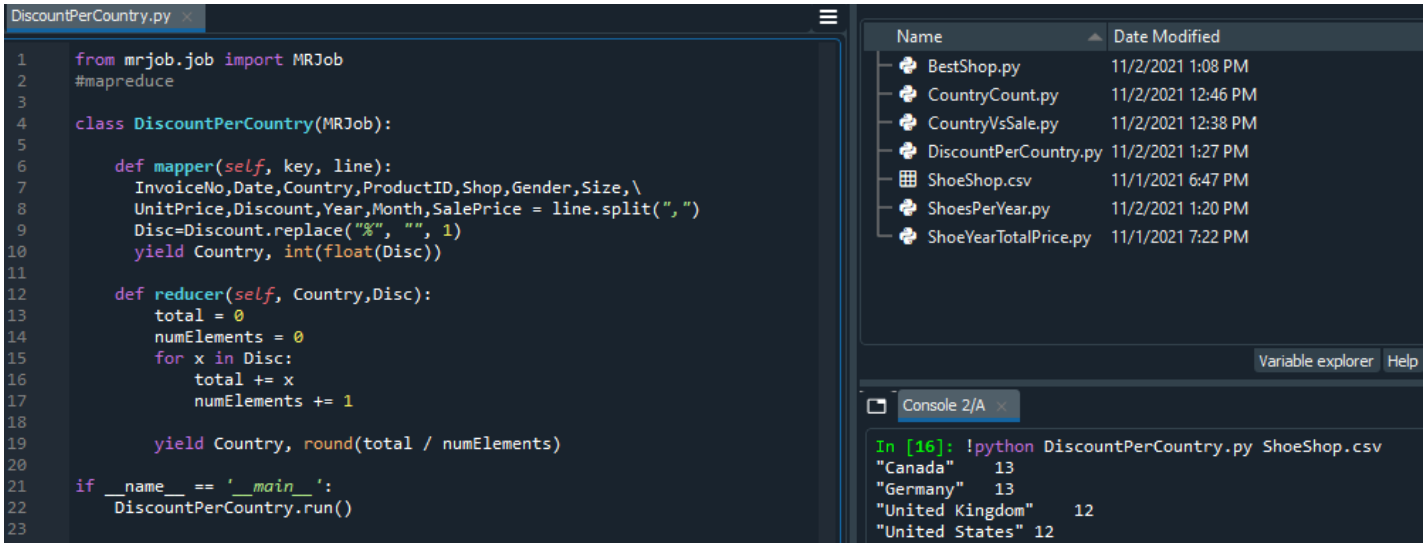
	InvoiceNo	Date	Country	ProductID	Shop	Gender	Size	UnitPrice	Discount	Year	Month	SalePrice
1.	52389	1/1/2014	United Kingdom	2152	UK2	Male	11	159	0%	2014	1	159
2.	52390	1/1/2014	United States	2230	US15	Male	11.5	199	20%	2014	1	159.2
3.	52391	1/1/2014	Canada	2160	CAN7	Male	9.5	149	20%	2014	1	119.2
...	---	---	---	---	---	---	---	---	---	---	---	---
14967	65777	12/31/2016	Germany	2156	GER1	Female	6.5	139	10%	2016	12	125.1

Mapping (Key=Country, Value= Discount)

United Kingdom, 0	United States, 20	Canada, 20	...	Germany, 10



Output (Spyder IDE)



Result

Average discount offered by each country is given below:

a) **Canada** has offered average **13 %** discount.
b) **Germany** has offered average **13 %** discount.
c) **United Kingdom** has offered average **12 %** discount.
d) **United States** has offered average **12 %** discount.

9. Store Dataset

Problem: What is the average tax paid by each City?

Solution:

i. Input the **Store** data set.

ii. Through **mapper** ( ) function, select the column **"City"** as Key and **"Payable Tax"** as value to each key as

(Key: value = City: Payable Tax).

iii. **Shuffle and sort** the key-value pair and then group them on the basis of key.

iv. Through **Reducer** ( ) function, calculate the average of **Tax** for each City and then yield the result in key-value pair where:

a. **Key** represents the **City**

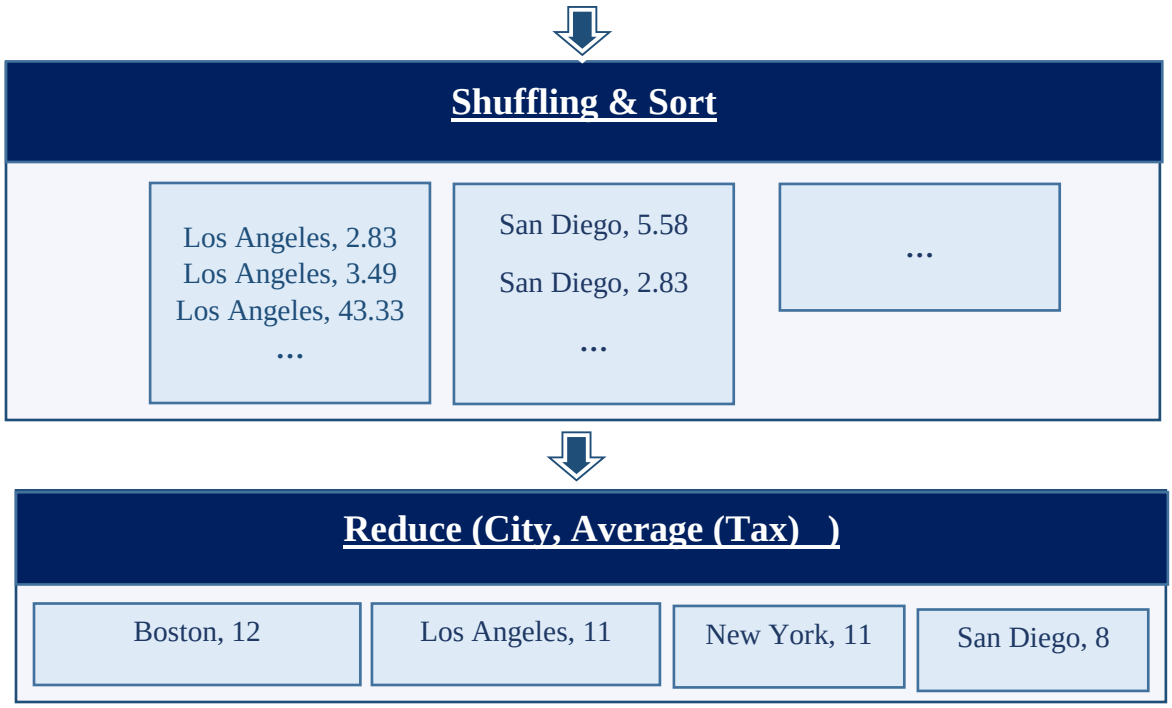
b. Value represents **Average Tax.**

Input Data

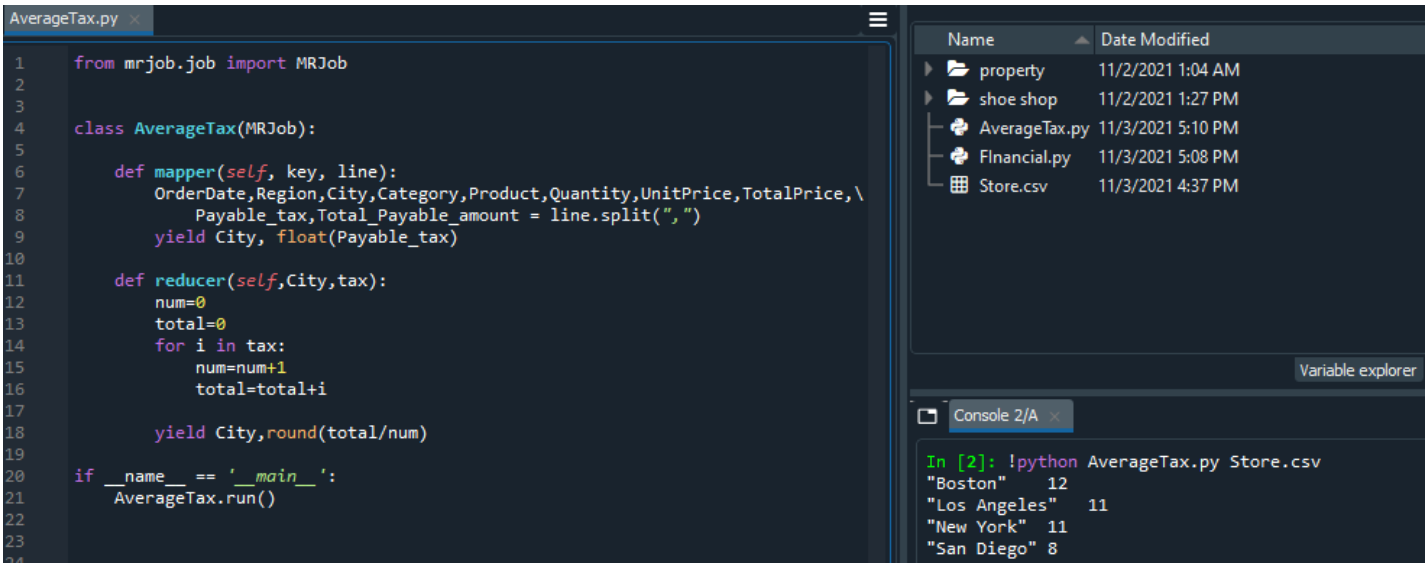
	OrderDate	Region	City	Category	Product	Quantity	UnitPrice	TotalPrice	Payable_tax	Total_Payable_amount
1.	43577	West	Los Angeles	Bars	Carrot	20	1.77	Rs 35.40	2.83	38.23
2.	43628	West	San Diego	Crackers	Whole Wheat	20	3.49	Rs 69.80	5.58	75.38
3.	43760	West	San Diego	Bars	Carrot	20	1.77	Rs 35.40	2.83	38.23
4.	44039	West	Los Angeles	Cookies	Arrowroot	20	2.18	Rs 43.60	3.49	47.09
...	...									
244	43637	West	Los Angeles	Bars	Carrot	306	1.77	Rs 541.62	43.33	584.95

Mapping (Key=City, Value= Payable Tax)

Los Angeles, 2.83	San Diego, 5.58	San Diego, 2.83	Los Angeles, 3.49	...	Los Angeles, 43.33



Output (Spyder IDE)



Result

The average tax paid by each City is given below:

a) **Boston** has paid average tax of **12.**
b) **Los Angeles** has paid average tax of **11.**
c) **New York** has paid average tax of **11.**
d) **San Diego** has paid average tax of **8.**

10. Store Dataset

Problem: How many Products have been sold in each Region?

Solution:

v. Input the **Store** data set.

vi. Through **mapper** ( ) function, select the column **"Region"** as Key and **"Quantity"** as value to each key as

(Key: value = Region: Quantity).

vii. **Shuffle and sort** the key-value pair and then group them on the basis of key.

viii. Through **Reducer** ( ) function, calculate the sum of **Quantity** for each Region and then yield the result in key-value pair where:

a. **Key** represents the **Region.**

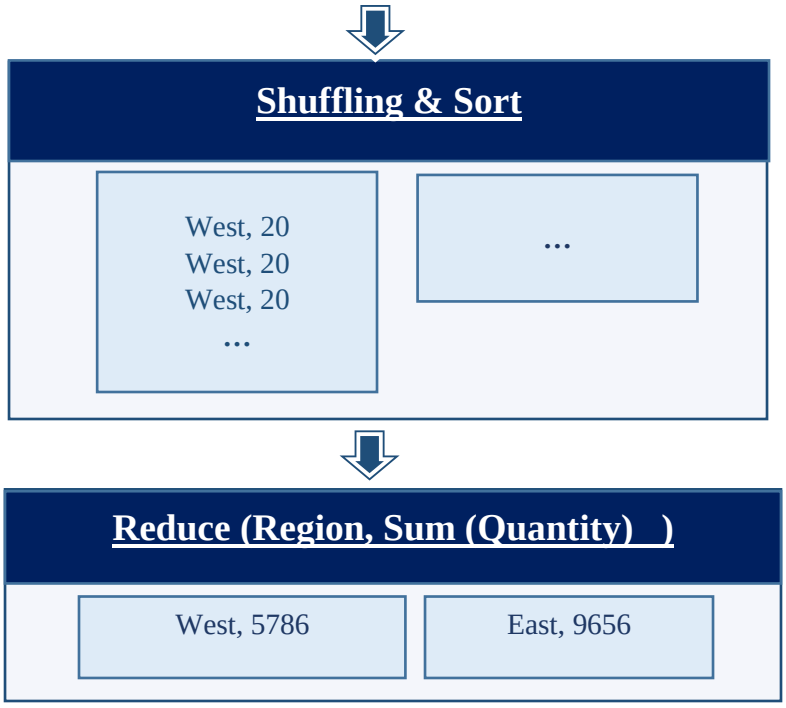
b. Value represents **total Quantity of products.**

Input Data

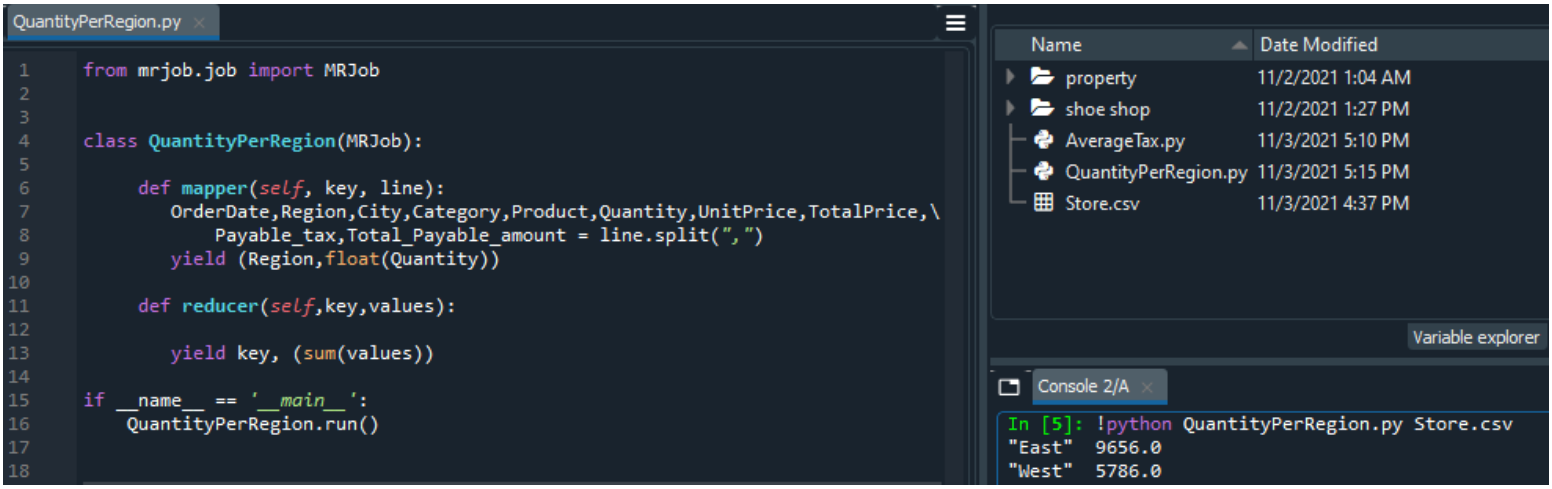
	OrderDate	Region	City	Category	Product	Quantity	UnitPrice	TotalPrice	Payable_tax	Total_Payable_amount
1.	43577	West	Los Angeles	Bars	Carrot	20	1.77	Rs 35.40	2.83	38.23
2.	43628	West	San Diego	Crackers	Whole Wheat	20	3.49	Rs 69.80	5.58	75.38
3.	43760	West	San Diego	Bars	Carrot	20	1.77	Rs 35.40	2.83	38.23
4.	44039	West	Los Angeles	Cookies	Arrowroot	20	2.18	Rs 43.60	3.49	47.09
...	…									
244	43637	West	Los Angeles	Bars	Carrot	306	1.77	Rs 541.62	43.33	584.95

Mapping (Key=Region, Value= Quantity)

West, 20	West, 20	West, 20	West, 20	…	West, 20



Output (Spyder IDE)



Result

The number of Products have been sold in each Region is given below:

a) In **East,** 9,656 products have been sold.
b) In **West,** 5,786 products have been sold.
