ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΑΣ, ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ
Τμήμα Πληροφορικής και Τηλεπικοινωνιών

BD-project1

Εργασία μαθήματος ΔΙΑΧΕΙΡΗΣΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

Συγγραφέας:
Λέκκας Γεώργιος, 2022201800108, dit18108@go.uop.gr
Κουνδουρόπουλος Αθανάσιος Συμεών, 2022201800097, dit18097@go.uop.gr

Δεκέμβριος 2021

Πίνακας περιεχομένων

1 Μελέτη των δεδομένων.....3
2 Εκπαιδευτικό υπόβαθρο πελατών.....5
3 Ανάδειξη των πελατών που είναι πιο πιθανό να αγοράσουν ένα νέο προϊόν κρασιού..7
4 Κατηγοριοποίηση των πελατών ανάλογα με την αγοραστική τους δυναμική.....10
5 Manual.....14

1 Μελέτη των δεδομένων

Nominal/Numeric

Nominal:

1. ID
2. Year_Birth
3. Education
4. Marital_Status
5. Dt_Customer
6. Complain
7. AcceptedCmp1
8. AcceptedCmp2
9. AcceptedCmp3
10. AcceptedCmp4
11. AcceptedCmp5
12. Response

Numeric:

1. Income
2. Kidhome
3. Teenhome
4. Recency
5. MntWines
6. MntFruits
7. MntMeatProducts
8. MntFishProducts
9. MntSweetProducts
10. MntGoldProds
11. NumDealsPurchases
12. NumWebPurchases
13. NumCatalogPurchases
14. NumStorePurchases
15. NumWebVisitsMonth

Outlier

Δεδομένο ότι ο μέσος όρος εισοδήματος στις δικές μας εγγραφές είναι 51687.46 ένα παράδειγμα outlier τιμής θα ήταν 1.500.000 καθώς είναι μια τιμή που αποκλίνει από τον μέσο όρο. Όσο αναφορά την ηλικία outliers μπορούν να θεωρηθούν ηλικίες <17 χρονών η >90 χρονών η ακόμα και ακραίες περιπτώσεις (πχ 250) που μπορεί να είναι λάθος

Προ-επεξεργασία

Η προ-επεξεργασία των δεδομένων που έγινε ήταν στο χαρακτηριστικό Income (Εισόδημα), καθώς σε κάποιες περιπτώσεις δεν υπήρχε κάποια τιμή και υπήρχε κενό, οπότε το αντικαταστήσαμε με την τιμή 0. Ο λόγος που κάναμε αυτή την επεξεργασία είναι ότι υπήρχε πρόβλημα όταν προσπαθήσαμε να επεξεργαστούμε την τιμή Income και να την μετατρέψουμε από String σε Integer. Επίσης, κάνουμε και τους απαραίτητους ελέγχους εγκυρότητας σε κάποια σημεία του κώδικα, για την αποφυγή κάποιου αντίστοιχου προβλήματος.

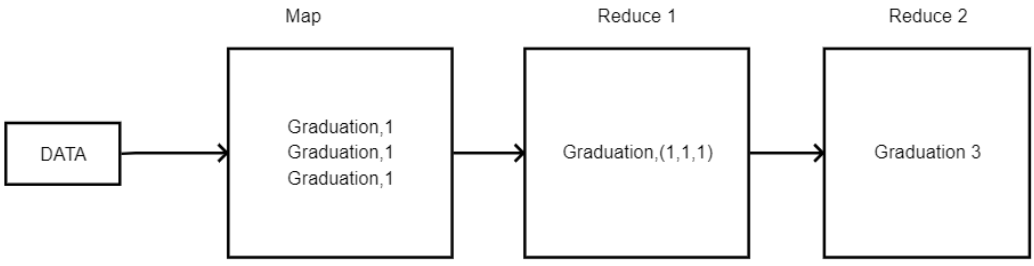
2 Εκπαιδευτικό υπόβαθρο πελατών

Εκπαιδευτικό υπόβαθρο πελατών {

Data = (5524, 1957, Graduation, Single, 58138, 0, 0, 04-09-2012, 58, 635, 88, 546, 172, 88, 88, 3, 8, 10, 4,7, 0, 0, 0, 0, 0, 0, 1)

```
void Map (key, String data values) {
        key = data(Graduation);
        emit(key, 1);
}

void Reduce (Text key, int [] counts) {
        int sum = 0;
        foreach (int i in counts) {
                sum += i;
        }
        emit(key, sum);
}
```

Παράδειγμα



Αποτέλεσμα εκτέλεσης του κώδικα

```
1 2n Cycle        203
2 Basic   54
3 Graduation      1127
4 Master  370
5 PhD     486
```

3 Ανάδειξη των πελατών που είναι πιο πιθανό να αγοράσουν ένα νέο προϊόν κρασιού

Ανάδειξη των πελατών που είναι πιο πιθανό να αγοράσουν ένα νέο προϊόν κρασιού {

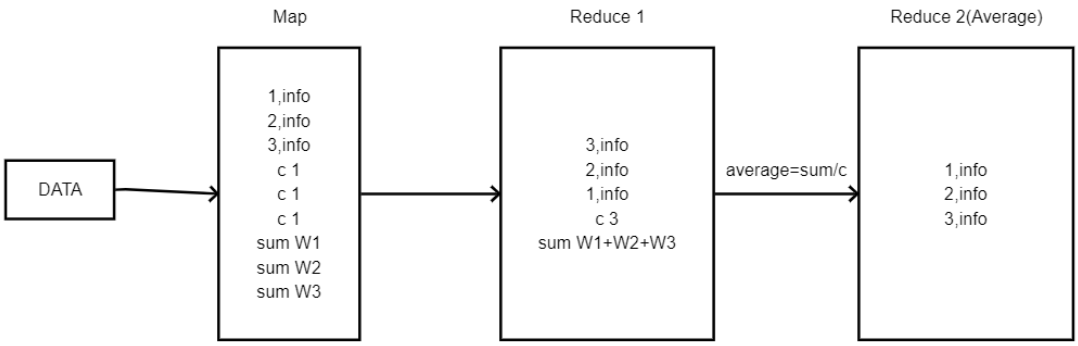
Data = (5524, 1957, Graduation, Single, 58138, 0, 0, 04-09-2012, 58, 635, 88, 546, 172, 88, 88, 3, 8, 10, 4,7, 0, 0, 0, 0, 0, 0, 1)

Info = (ID, age, Education, Marital_Status, Income, MntWines)

```
void Map (key, String data values) {
        int i=0
        Text c
        Text sum
        Info = (ID, age, Education, Marital_Status, Income, MntWines)
        emit(c, 1);
        emit(Text(i++), Info);
        emit(sum, MntWines );
}

void Reduce1 (Text key, Text Values) {
        int sum = 0;
        Text textvalue;

        foreach (Text value in values) {
                if(key==sum) {
                        sum += value;
                } else if(key==c) {
                        c++;
                } else {
                        textvalue = value;
                }
        }
                emit(Text(sum), sum);
                emit(count,c);
                emit(key,textvalue);
}
```

```
void Reduce2(Text key, Text Values) {
        int sum = 0;
        int count = 0;
        int average = 0;
        ArrayList <MntWinesObject> list = new ArrayList();

        foreach (Text value in values) {
                if(key!=sum && key!= count) {
                        info = value;
                        list.add(new MntWinesObject(info));
                        list.sort;
                } else
                   if(key==c)
                        count = c;
                    if(key==sum)
                        sum = sum;

        }
        average = sum/count;
        if(MntWinesObject.MntWines>avergage) {
                emit(i, MntWinesObject.info);
        }
}
```

Παράδειγμα για 3 εγγραφές



Αποτέλεσμα εκτέλεσης του κώδικα

1	1	737 72 Graduation Married 80360 1493
2	2	3174 62 Graduation Together 87771 1492
3	3	5536 62 Graduation Together 87771 1492
4	4	1103 45 Graduation Married 81929 1486
5	5	5547 39 Graduation Married 84169 1478
6	6	8362 39 Graduation Married 84169 1478
7	7	3009 59 Graduation Widow 71670 1462
8	8	1665 57 Graduation Divorced 64140 1459
9	9	9743 66 Graduation Married 76998 1449
10	10	11088 50 Graduation Together 78642 1396
11	11	4580 52 Graduation Married 75759 1394
12	12	4943 68 Graduation Married 70503 1379
13	13	9260 76 Graduation Married 70356 1349
14	14	7431 30 Graduation Single 68126 1332
15	15	3138 65 Graduation Single 91249 1324
16	16	4475 72 Graduation Married 69098 1315
17	17	6292 35 Graduation Married 82333 1311
18	18	10140 38 Graduation Together 70123 1308

4 Κατηγοριοποίηση των πελατών ανάλογα με την αγοραστική τους δυναμική

Κατηγοριοποίηση των πελατών ανάλογα με την αγοραστική τους δυναμική{

Data = (5524, 1957, Graduation, Single, 58138, 0, 0, 04-09-2012, 58, 635, 88, 546, 172, 88, 88, 3, 8, 10, 4,7, 0, 0, 0, 0, 0, 0, 1)

Info = (ID, Income, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds)

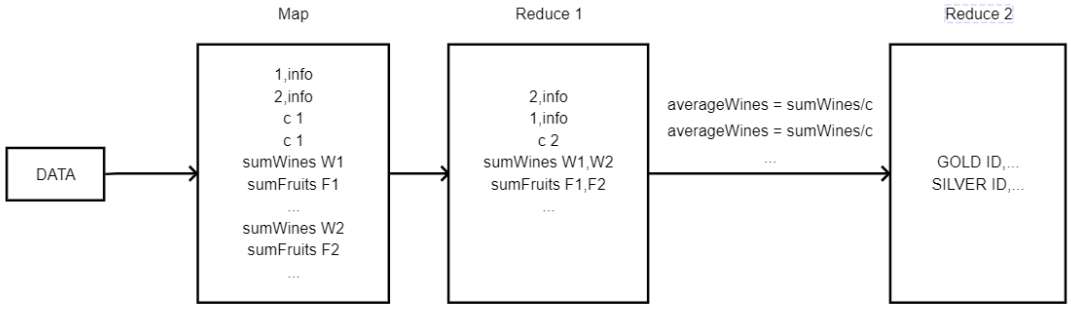
```
void Map(key, String data values) {
        int i = 0;
        Text c;
        Text sumWines;
        Text sumFruits;
        ...
        Info = (ID, Income, MntWines, MntFruits, MntMeatProducts, MntFishProducts,
        MntSweetProducts, MntGoldProds)
        emit(c, 1);
        emit(Text(i++), Info);
        emit(Text(sumWines), MntWines );
        emit(Text(sumFruits), MntFruits );
        ...
}
```

```
void Reduce1(Text key, Text Values) {
 Text sumWines;
 Text sumFruits;
 ...
Text textvalue
        foreach (Text value in values) {
                if(key==sumWines || key==sumFruits ...) {
                        sum += value;
                } else if(key==c) {
                        c++;
                } else {
                        textvalue = value;
                }
        }
        emit(Text(sumWines), sumWines );
        emit(Text(sumFruits),sumFruits );
        ...
        emit(count, c);
        emit(key, textvalue);
}
```

```
void Reduce2(Text key, Text Values) {
        Text sumWines;
        Text sumFruits;
        ...
        int count = 0;
        int averageWines = 0;
        int averageFruits = 0;
        ...
        ArrayList <Customers> gold = new ArrayList();
        ArrayList <Customers> silver = new ArrayList();
        foreach (Text value in values) {
                if(key!=sumWines && key!=sumFruits ...) {
                        info = value;
                        new Costumer(info);
                } else {
                        if(key==c) {
                                count = c;
                        }
                        if(key==sumWines || key==sumFruits ...) {
                                sum+=sum;
                                ...
                        }
        }
        averageWines=sumWines/count;
        averageFruits=sumFruits/count;
        if(Costumer.year==21 && income>69500 && mntwines>averageWines ...)
                gold.add(costumer);
                ...
                silver.add(costumer);
        foreach (Costumer in gold) {
                emit(Gold, Costumer.info, id);
        }
```

```
        foreach (Costumer in silver) {
                emit(Silver, Costumer.info, id);
        }
}
```

Παράδειγμα για 2 εγγραφές



Αποτέλεσμα εκτέλεσης του κώδικα

```
1 Gold    1225 1577 2186 2963 3005 3010 3334 4676 4767 5350 5735 5848 7010 7059 7723 10129
2 Silver  569 590 1150 1215 1232 1544 1553 2147 2254 2345 2429 3139 3698 4597 5453 5831 6055
```

5 Manual

Η υλοποίηση του Project έγινε σε Linux Ubuntu και με την χρήση του Eclipse. Ή έκδοση του Hadoop που χρησιμοποιήσαμε ήταν η 3.2.1 και η έκδοση της java 8. Επίσης, έγινε και η χρήση των παρακάτω libraries:

- commons-cli-1.2.jar
- hadoop-common-3.2.1.jar
- hadoop-mapreduce-client-core-3.2.1.jar

Τέλος η εκτέλεση του Project γίνεται με τις παρακάτω εντολές:

- hadoop jar /home/hdoop/testing.jar Main /editeddata/personality_analysis.csv /r_output