Bigdata Systems - Assignment 1 (S1-22_SEZG522)

Submitted by Group 11

- Noel John K - 2021MT93693
- Pavithra S – 2021MT93542
- Jayanthi Sangita M - 2021MT93337

1 Dataset

Dataset source - *https://www.kaggle.com/datasets/thedevastator/chemicals-in-cosmetics-what-s-really-in-your?resource=download*

This dataset is provided by Kaggle, and it contains ***114,297*** records of information on the chemicals used in cosmetics, including the name of the chemicals, the company that manufactures it, the primary category it is used in, and the date it was first reported.

The dataset contains following attributes:

index,CDPHId,ProductName,CSFId,CSF,CompanyId,CompanyName,BrandName,PrimaryCategoryId,PrimaryCategory,SubCategoryId,SubCategory,CasId,CasNumber,ChemicalId,ChemicalName,InitialDateReported,MostRecentDateReported,DiscontinuedDate,ChemicalCreatedAt,ChemicalUpdatedAt,ChemicalDateRemoved,ChemicalCount

2 Assumptions

The following assumptions are made so that complex logics can be avoided.

- Values containing commas are neglected since the comma is considered as a delimiter.
- Instead of using ids as keys, we are taking name as the primary key. Because only few fields had ids.

3 Hadoop Cluster

We were facing issues with BITS remote labs. So, we have spin up a Virtual machine in Azure and followed an article to setup a single node Hadoop Cluster. The reference articles are added to the reference section of this document.

4 Execution

The MapReduce jobs can be executed in the cluster we have setup by executing the following command.

hadoop jar libs/hadoop-streaming-3.3.4.jar -files mapper.py,reducer.py,chemicals-in-cosmetics-3.csv -mapper mapper.py -reducer reducer.py -input chemicals-in-cosmetics-3.csv -output output

In order to test the application locally, the following command can be used.

cat .\chemicals-in-cosmetics-3.csv | python .\mapper.py | python .\reducer.py

5 Analysis 1

5.1 Analysis Performed

Finding the unique cosmetic products launched by a company - In this MapReduce program, we must find out the unique products launched by a company irrelevant of its brand name. This will help to identify how much cosmetic products are patented to each company.

5.2 Input & Output Attributes

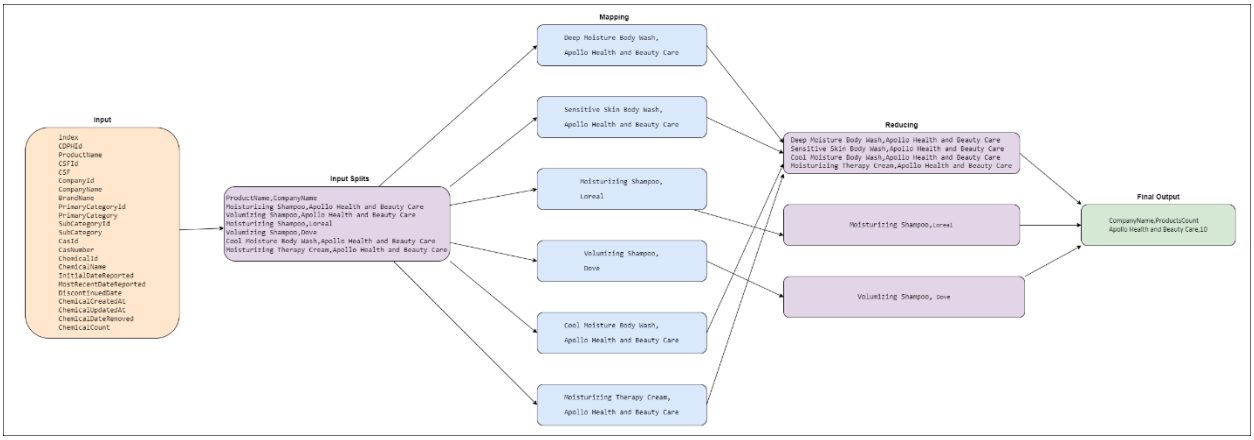
5.2.1 Input Attributes

CompanyName, *ProductName*

5.2.2 Output Attributes

CompanyName, *ProductNameCount*

5.3 MapReduce Diagrams



5.4 Mapper & Reducer Pseudo Codes

5.4.1 Mapper Pseudo Code

```
class Mapper:
        method map(fullColumns):
                columns = fullColumns.split(‘,’)
                if(length(columns) == total_colums)
                        companyName = columns[companyNamePosition]
                        productName = columns[productNamePosition]
                        write(selected columns)
```

5.4.2 Reducer Pseudo Code

```
class Reducer:
        method reduce(companyName,  productName)
                uniqueProducts = companyName productName #uniqueProducts is a set

                key = companyName
                value = count(companyName, productName)
                dict = {key:value} #key is string, value is a counter where unique products are stored

                if key not in dict:
                        dict.getValue(key).add(1)
                else:
                        dict.getValue(key)+1
                write(dict.key, dict.value)
```

5.5 Mapper & Reducer Programs

5.5.1 Mapper Program

```
#!/usr/bin/env python3
import sys

delimiter = ","

def map():
   for line in sys.stdin:
      rows = line.strip()
      columns = rows.split(delimiter)

      if len(columns) == 23:
         product_name = columns[2]
         company_name = columns[6]

         print(f"{company_name}{delimiter}{product_name}")


if __name__ == "__main__":
   map()
```

5.5.2 Reducer Program

```
#!/usr/bin/env python3
import sys

delimiter = ","
total_products = set()
unique_products = {}


def reduce():
    my_iterator = iter(sys.stdin.readline, "")
    header = next(my_iterator)
    company_name_header, product_name_header = header.strip().split(delimiter)
    print(f"{company_name_header}{delimiter}ProductsCount")

    for line in sys.stdin:
        line = line.strip()

        company_name, product_name = line.split(delimiter)
        key = f"{company_name}{delimiter}{product_name}"
        total_products.add(key)

    for key in total_products:
        company_name, product_name = key.split(delimiter)

        if company_name in unique_products.keys():
            count = unique_products[company_name]
            unique_products[company_name] = count + 1
        else:
            unique_products[company_name] = 1

    for company_name in unique_products.keys():
        print(f"{company_name}{delimiter}{unique_products[company_name]}")


if __name__ == "__main__":
    reduce()
```

5.6 Statistics

```
2022-10-31 06:30:56,504 INFO mapred.Task: Final Counters for attempt_local610285712_0001_m_000000_0: Counters: 17
        File System Counters
                FILE: Number of bytes read=56044226
                FILE: Number of bytes written=32672356
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=114299
                Map output records=63743
                Map output bytes=3591711
                Map output materialized bytes=3719469
                Input split bytes=96
                Combine input records=0
                Spilled Records=63743
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=65
                Total committed heap usage (bytes)=240123904
        File Input Format Counters
                Bytes Read=27950723
2022-10-31 06:30:56,508 INFO mapred.LocalJobRunner: Finishing task: attempt_local610285712_0001_m_000000_0
```

```
2022-10-31 06:30:57,055 INFO mapred.Task: Final Counters for attempt_local610285712_0001_r_000000_0: Counters: 24
        File System Counters
                FILE: Number of bytes read=63483196
                FILE: Number of bytes written=36400900
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=21626
                Reduce shuffle bytes=3719469
                Reduce input records=63743
                Reduce output records=385
                Spilled Records=63743
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=240123904
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=9075
2022-10-31 06:30:57,058 INFO mapred.LocalJobRunner: Finishing task: attempt_local610285712_0001_r_000000_0
2022-10-31 06:30:57,058 INFO mapred.LocalJobRunner: reduce task executor complete.
```

```
2022-10-31 06:30:57,680 INFO mapreduce.Job: Job job_local610285712_0001 completed successfully
2022-10-31 06:30:57,690 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=119527422
                FILE: Number of bytes written=69073256
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=114299
                Map output records=63743
                Map output bytes=3591711
                Map output materialized bytes=3719469
                Input split bytes=96
                Combine input records=0
                Combine output records=0
                Reduce input groups=21626
                Reduce shuffle bytes=3719469
                Reduce input records=63743
                Reduce output records=385
                Spilled Records=127486
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=65
                Total committed heap usage (bytes)=480247808
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=27950723
        File Output Format Counters
                Bytes Written=9075
2022-10-31 06:30:57,690 INFO streaming.StreamJob: Output directory: output
```

5.7 Sample Input & Output data's

5.7.1 Input data

index,CDPHId,ProductName,CSFId,CSF,CompanyId,CompanyName,BrandName,PrimaryCategoryId,PrimaryCategory,SubCategoryId,SubCategory,CasId,CasNumber,ChemicalId,ChemicalName,InitialDateReported,MostRecentDateReported,DiscontinuedDate,ChemicalCreatedAt,ChemicalUpdatedAt,ChemicalDateRemoved,ChemicalCount

33518,11448,Deep Moisture Body Wash,,,475,Apollo Health and Beauty Care,Equate,6,Bath Products,159,Body Washes and Soaps,656,13463-67-7,16729,Titanium dioxide,05/20/2010,07-01-10,,05/20/2010,05/20/2010,,1

40705,14452,COOL MOISTURE BODY WASH,,,475,Apollo Health and Beauty Care,Equate,6,Bath Products,159,Body Washes and Soaps,656,13463-67-7,22125,Titanium dioxide,07-01-10,07-01-10,,07-01-10,07-01-10,,1

40708,14454,Sensitive Skin Body Wash,,,475,Apollo Health and Beauty Care,Equate,6,Bath Products,159,Body Washes and Soaps,656,13463-67-7,22127,Titanium dioxide,07-01-10,07-01-10,,07-01-10,07-01-10,,1

40714,14458,Moisturizing Therapy Cream,,,475,Apollo Health and Beauty Care,Natural Concepts,90,Skin Care Products ,102,Skin Moisturizers (making a cosmetic claim),656,13463-67-7,22131,Titanium dioxide,07-01-10,07-01-10,,07-01-10,07-01-10,,1

41452,14656,Tropical Renewal Softening Body Wash,,,475,Apollo Health and Beauty Care,Equate,6,Bath Products,159,Body Washes and Soaps,656,13463-67-7,22408,Titanium dioxide,07-12-10,07-12-10,,07-12-10,07-12-10,,1

43022,15133,Frizz Release Hold Gel,,,475,Apollo Health and Beauty Care,Natural Concepts,18,Hair Care Products (non-coloring),26,Hair Styling Products,656,13463-67-7,23224,Titanium dioxide,08/20/2010,08/20/2010,,08/20/2010,08/20/2010,,1

48508,16825,Cool Moisture Body Wash,,,475,Apollo Health and Beauty Care,IMAGE ESSENTIALS,6,Bath Products,159,Body Washes and Soaps,656,13463-67-7,26185,Titanium dioxide,07-05-11,07-05-11,,07-05-11,07-05-11,,1

48509,16826,Deep Moisture Boday Wash,,,475,Apollo Health and Beauty Care,IMAGE ESSENTIALS,6,Bath Products,159,Body Washes and Soaps,656,13463-67-7,26186,Titanium dioxide,07-05-11,07-05-11,,07-05-11,07-05-11,,1

48677,16912,Moisturizing Shampoo,,,475,Apollo Health and Beauty Care,Rusk,18,Hair Care Products (non-coloring),25,Hair Shampoos (making a cosmetic claim) ,656,13463-67-7,26306,Titanium dioxide,08-05-11,08-05-11,,08-05-11,08-05-11,,1

48678,16915,Volumizing Shampoo,,,475,Apollo Health and Beauty Care,Rusk,18,Hair Care Products (non-coloring),25,Hair Shampoos (making a cosmetic claim) ,656,13463-67-7,26308,Titanium dioxide,08-05-11,08-05-11,,08-05-11,08-05-11,,1

5.7.2 Output data

CompanyName,ProductsCount
Apollo Health and Beauty Care,10

6 Analysis 2

6.1 Analysis Performed

Finding all the chemicals associated with a product launched by a company under one brand name – Here the output of the MapReduce program will give us the chemicals used to manufacture the cosmetic product launched by a company under a brand name.

6.2 Input & Output Attributes

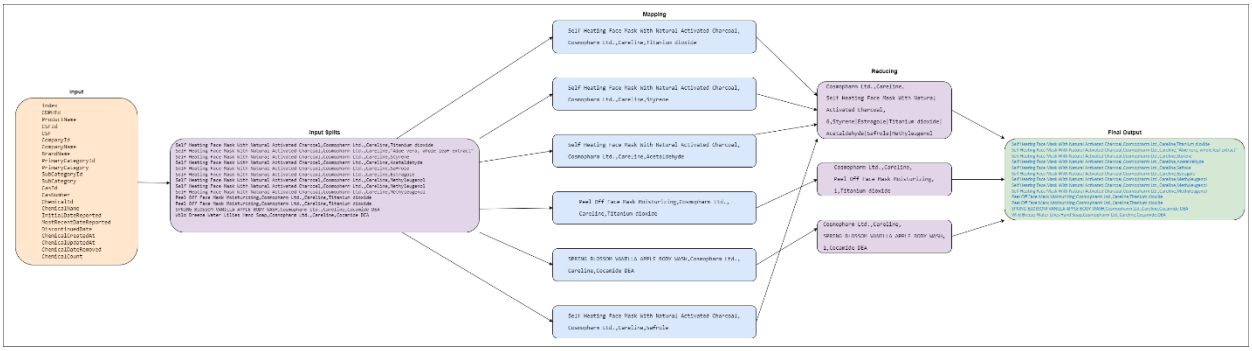
6.2.1 Input attribute

CompanyName, *BrandName*, *ProductName*, *ChemicalName*

6.2.2 Output Attribute

CompanyName, *BrandName*, *ProductName*, *ChemicalCount*, *ChemicalName*

6.3 MapReduce Diagrams



6.4 Mapper & Reducer Pseudo Codes

6.4.1 Mapper Pseudo Code

```
class Mapper:
        method map(fullColumns):
                columns = fullColumns.split(',')
                if(length(columns) == total_colums)
                        companyName = columns[companyNamePosition]
                        brandName = columns[brandNamePosition]
                        productName = columns[productNamePosition]
                        chemicalName = columns[chemicalNamePosition]
                        write(selected columns)
```

6.4.2 Reducer Pseudo Code

```
class Reducer:
        method reduce(companyName, brandName, productName, chemicalName)
                key = companyName, brandName, productName
                value = chemicalName
                dict = {key:value} # key is string, value is a set
                if value not in dict:
                        dict.getValue().add(value)
                        write(dict.key, dict.value)
```

6.5 Mapper & Reducer Programs

6.5.1 Mapper Program

```
#!/usr/bin/env python3
import sys

delimiter = ","


def map():
   for line in sys.stdin:
      rows = line.strip()
      columns = rows.split(delimiter)

      if len(columns) == 23:
         company_name = columns[6]
         brand_name = columns[7]
         product_name = columns[2]
         chemical_name = columns[15]


print(f"{company_name}{delimiter}{brand_name}{delimiter}{product_name}{delimiter}{chemical_name}")


if __name__ == "__main__":
   map()
```

6.5.2 Reducer Program

```
#!/usr/bin/env python3
import sys

delimiter = ","
content_separator = "|"
unique_products = {}


def reduce():
    my_iterator = iter(sys.stdin.readline, "")
    header = next(my_iterator)
    company_name_header, brand_name_header, product_name_header, chemical_name_header =
header.strip().split(delimiter)

print(f"{company_name_header}{delimiter}{brand_name_header}{delimiter}{product_name_header}{deli
miter}ChemicalCount{delimiter}{chemical_name_header}")

    for line in sys.stdin:
        line = line.strip()

        company_name, brand_name, product_name, chemical_name = line.split(delimiter)
        key = f"{company_name}{delimiter}{brand_name}{delimiter}{product_name}"

        if key in unique_products.keys():
            unique_chemicals = unique_products[key]
            unique_chemicals.add(chemical_name)
        else:
            unique_chemicals = set()
            unique_chemicals.add(chemical_name)
            unique_products[key] = unique_chemicals

    for key in unique_products.keys():

print(f"{key}{delimiter}{len(unique_products[key])}{delimiter}{content_separator.join(unique_products[key
])}")


if __name__ == "__main__":
    reduce()
```

6.6 Statistics

```
2022-10-31 06:16:41,777 INFO mapred.Task: Task:attempt_local1727250169_0001_m_000000_0 is done. And is in the process of committing
2022-10-31 06:16:41,779 INFO mapred.LocalJobRunner: Records R/W=554/1
2022-10-31 06:16:41,779 INFO mapred.Task: Task 'attempt_local1727250169_0001_m_000000_0' done.
2022-10-31 06:16:41,787 INFO mapred.Task: Final Counters for attempt_local1727250169_0001_m_000000_0: Counters: 17
        File System Counters
                FILE: Number of bytes read=56043957
                FILE: Number of bytes written=34497624
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=114299
                Map output records=63743
                Map output bytes=5411817
                Map output materialized bytes=5541911
                Input split bytes=96
                Combine input records=0
                Spilled Records=63743
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=227
                Total committed heap usage (bytes)=470286336
        File Input Format Counters
                Bytes Read=27950723
2022-10-31 06:16:41,787 INFO mapred.LocalJobRunner: Finishing task: attempt_local1727250169_0001_m_000000_0
2022-10-31 06:16:41,787 INFO mapred.LocalJobRunner: map task executor complete.
2022-10-31 06:16:41,793 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2022-10-31 06:16:41,794 INFO mapred.LocalJobRunner: Starting task: attempt_local1727250169_0001_r_000000_0
2022-10-31 06:16:41,801 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-10-31 06:16:41,801 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failu
es: false
```

```
2022-10-31 06:16:42,468 INFO mapred.Task: Final Counters for attempt_local1727250169_0001_r_000000_0: Counters: 24
        File System Counters
                FILE: Number of bytes read=67127811
                FILE: Number of bytes written=42044895
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=24043
                Reduce shuffle bytes=5541911
                Reduce input records=63743
                Reduce output records=21956
                Spilled Records=63743
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=470286336
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=2005360
2022-10-31 06:16:42,469 INFO mapred.LocalJobRunner: Finishing task: attempt_local1727250169_0001_r_000000_0
```

```
2022-10-31 06:16:42,996 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=123171768
                FILE: Number of bytes written=76542519
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=114299
                Map output records=63743
                Map output bytes=5411817
                Map output materialized bytes=5541911
                Input split bytes=96
                Combine input records=0
                Combine output records=0
                Reduce input groups=24043
                Reduce shuffle bytes=5541911
                Reduce input records=63743
                Reduce output records=21956
                Spilled Records=127486
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=227
                Total committed heap usage (bytes)=940572672
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=27950723
        File Output Format Counters
                Bytes Written=2005360
2022-10-31 06:16:42,997 INFO streaming.StreamJob: Output directory: output
```

6.7 Sample Input and Output data's

6.7.1 Input data

index,CDPHId,ProductName,CSFId,CSF,CompanyId,CompanyName,BrandName,PrimaryCategoryId,PrimaryCategory,SubCategoryId,SubCategory,CasId,CasNumber,ChemicalId,ChemicalName,InitialDateReported,MostRecentDateReported,DiscontinuedDate,ChemicalCreatedAt,ChemicalUpdatedAt,ChemicalDateRemoved,ChemicalCount
113768,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products,93,Skin Cleansers,656,13463-67-7,67675,Titanium dioxide,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,03/20/2020,6
113769,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products,93,Skin Cleansers,1108,,67676,"Aloe vera, whole leaf extract",03/20/2020,03/20/2020,,03/20/2020,03/20/2020,,6
113770,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products,93,Skin Cleansers,620,100-42-5,67677,Styrene,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,,6
113771,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products,93,Skin Cleansers,2,75-07-0,67678,Acetaldehyde,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,,6
113772,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products,93,Skin Cleansers,608,94-59-7,67679,Safrole,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,,6
113773,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products,93,Skin Cleansers,293,140-67-0,67680,Estragole,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,,6
113774,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products,93,Skin Cleansers,442,93-15-2,67681,Methyleugenol,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,03/20/2020,6
113775,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products,93,Skin Cleansers,442,93-15-2,67682,Methyleugenol,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,,6
113776,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products,93,Skin Cleansers,442,93-15-2,67683,Methyleugenol,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,03/20/2020,6
113777,41309,Peel Off Face Mask Moisturizing,,,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products ,95,Facial Masks,656,13463-67-7,67684,Titanium dioxide,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,,1
113778,41309,Peel Off Face Mask Moisturizing,,,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products ,102,Skin Moisturizers (making a cosmetic claim),656,13463-67-7,67684,Titanium dioxide,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,,1
114295,41449,SPRING BLOSSOM VANILLA APPLE BODY WASH,,,1388,Cosmopharm Ltd.,Careline,6,Bath Products,159,Body Washes and Soaps,969,,67905,Cocamide DEA,04/30/2020,04/30/2020,,04/30/2020,04/30/2020,,1
114296,41450,Wild Breeze Water Lilies Hand Soap,,,1388,Cosmopharm Ltd.,Careline,6,Bath Products,159,Body Washes and Soaps,969,,67906,Cocamide DEA,04/30/2020,04/30/2020,,04/30/2020,04/30/2020,,1

6.7.2 Output data

CompanyName,BrandName,ProductName,ChemicalCount,ChemicalName
Cosmopharm Ltd.,Careline,Self Heating Face Mask With Natural Activated Charcoal,6,Styrene|Estragole|Titanium dioxide|Acetaldehyde|Safrole|Methyleugenol
Cosmopharm Ltd.,Careline,Peel Off Face Mask Moisturizing,1,Titanium dioxide
Cosmopharm Ltd.,Careline,SPRING BLOSSOM VANILLA APPLE BODY WASH,1,Cocamide DEA
Cosmopharm Ltd.,Careline,Wild Breeze Water Lilies Hand Soap,1,Cocamide DEA

7 Analysis 3

7.1 Analysis Performed

Finding primary category of cosmetics which has highest discontinued chemicals – In this MapReduce problem, the output will give us the primary category which contains the highest no. of chemicals that are discontinued.

7.2 Input & Output Attributers

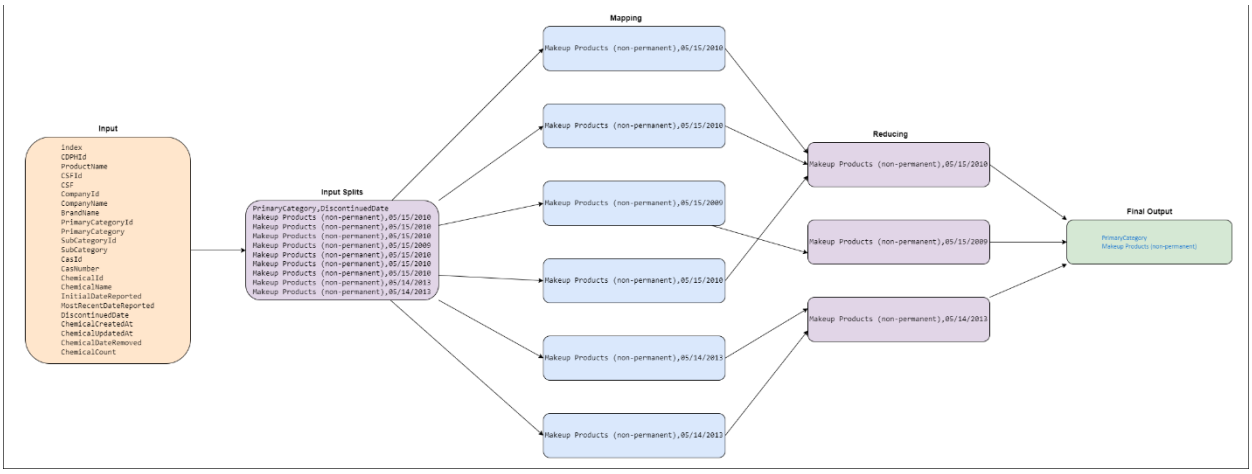
7.2.1 Input Attribute

PrimaryCategory, *ChemicalDateRemoved*

7.2.2 Output Attribute

PrimaryCategory

7.3 MapReduce Diagrams



7.4 Mapper & Reducer Pseudo Codes

7.4.1 Mapper Pseudo Code

```
class Mapper:
        method map(fullColumns):
                columns = fullColumns.split(‘,’)
                if(length(columns) == total_colums)
                        primaryCategory = columns[primaryCategoryPosition]
                        chemicalDateRemoved = columns[chemicalDateRemovedPosition]
                        write(selected columns)
```

7.4.2 Reducer Pseudo Code

```
class Reducer:
        method reduce(primaryCategory, chemicalDateRemoved)
                key = primaryCategory
                value = chemicalName
                dict = {key:value} # key is string, value is a counter

                if key not in dict:
                        dict.getValue(key).add(1)
                else:
                        dict.getValue(key)+1

                write(dict.key, dict.value)
```

7.5 Mapper & Reducer Programs

7.5.1 Mapper Program

```
#!/usr/bin/env python3
import sys

delimiter = ","

def map():
    for line in sys.stdin:
        rows = line.strip()
        columns = rows.split(delimiter)

        if len(columns) == 23:
            primary_category = columns[9]
            discontinued_date = columns[18]

            print(f"{primary_category}{delimiter}{discontinued_date}")

if __name__ == "__main__":
    map()
```

7.5.2 Reducer Program

```
#!/usr/bin/env python3
import sys
import collections

delimiter = ","
primary_category_with_discontinued_chemicals = collections.Counter()

def reduce():
    my_iterator = iter(sys.stdin.readline, "")
    header = next(my_iterator)
    primary_category_header, discontinued_date_header = header.strip().split(delimiter)
    print(f"{primary_category_header}")

    for line in sys.stdin:
        line = line.strip()

        primary_category, discontinued_date = line.split(delimiter)

        if discontinued_date is not None and discontinued_date!= "":
            if primary_category in primary_category_with_discontinued_chemicals.keys():
                count = primary_category_with_discontinued_chemicals[primary_category]
                primary_category_with_discontinued_chemicals[primary_category] = count + 1
            else:
                primary_category_with_discontinued_chemicals[primary_category] = 1

    print(primary_category_with_discontinued_chemicals.most_common(1)[0][0])


if __name__ == "__main__":
    reduce()
```

7.6 Statistics

```
2022-10-31 06:50:09,520 INFO mapred.Task: Final Counters for attempt_local9330798_0001_m_000000_0: Counters: 17
        File System Counters
                FILE: Number of bytes read=56044275
                FILE: Number of bytes written=30985573
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=114299
                Map output records=63743
                Map output bytes=1911355
                Map output materialized bytes=2038847
                Input split bytes=96
                Combine input records=0
                Spilled Records=63743
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=251
                Total committed heap usage (bytes)=478150656
        File Input Format Counters
                Bytes Read=27950723
2022-10-31 06:50:09,524 INFO mapred.LocalJobRunner: Finishing task: attempt_local9330798_0001_m_000000_0
```

```
2022-10-31 06:50:09,989 INFO mapred.Task: Task 'attempt_local9330798_0001_r_000000_0' done.
2022-10-31 06:50:09,990 INFO mapred.Task: Final Counters for attempt_local9330798_0001_r_000000_0: Counters: 24
        File System Counters
                FILE: Number of bytes read=60122001
                FILE: Number of bytes written=33024480
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=1236
                Reduce shuffle bytes=2038847
                Reduce input records=63743
                Reduce output records=2
                Spilled Records=63743
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=478150656
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=60
2022-10-31 06:50:09,990 INFO mapred.LocalJobRunner: Finishing task: attempt_local9330798_0001_r_000000_0
2022-10-31 06:50:09,990 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-10-31 06:50:10,010 INFO mapreduce.Job:  map 100% reduce 100%
2022-10-31 06:50:10,010 INFO mapreduce.Job: Job job_local9330798_0001 completed successfully
```

```
2022-10-31 06:50:10,010 INFO mapreduce.Job: Job job_local9330798_0001 completed successfully
2022-10-31 06:50:10,019 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=116166276
                FILE: Number of bytes written=64010053
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=114299
                Map output records=63743
                Map output bytes=1911355
                Map output materialized bytes=2038847
                Input split bytes=96
                Combine input records=0
                Combine output records=0
                Reduce input groups=1236
                Reduce shuffle bytes=2038847
                Reduce input records=63743
                Reduce output records=2
                Spilled Records=127486
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=251
                Total committed heap usage (bytes)=956301312
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=27950723
        File Output Format Counters
                Bytes Written=60
2022-10-31 06:50:10,019 INFO streaming.StreamJob: Output directory: output
```

7.7 Sample Input & Output data's

7.7.1 Input data

index,CDPHId,ProductName,CSFId,CSF,CompanyId,CompanyName,BrandName,PrimaryCategoryId,PrimaryCategory,SubCategoryId,SubCategory,CasId,CasNumber,ChemicalId,ChemicalName,InitialDateReported,MostRecentDateReported,DiscontinuedDate,ChemicalCreatedAt,ChemicalUpdatedAt,ChemicalDateRemoved,ChemicalCount

416,254,COLOR TREND LIQUID EYE LINER BRIGHTS-ALL SHADES ,,,4,New Avon LLC,AVON,44,Makeup Products (non-permanent),46,Eyeliner/Eyebrow Pencils,656,13463-67-7,265,Titanium dioxide,09-01-09,08/28/2013,05/15/2010,09-01-09,09-01-09,,14572,1359,AVON SHIMMER SWIRLS FACE ILLUMINATOR-ALL SHADES ,,,4,New Avon LLC,AVON,44,Makeup Products (non-permanent),49,Face Powders,656,13463-67-7,1488,Titanium dioxide,09/21/2009,08/28/2013,05/15/2010,09/21/2009,09/21/2009,,14733,1439,AVON 8-IN-1! LIP PALETTE-ALL SHADES ,,,4,New Avon LLC,AVON,44,Makeup Products (non-permanent),52,Lip Gloss/Shine,656,13463-67-7,1576,Titanium dioxide,09/22/2009,08/28/2013,05/15/2010,09/22/2009,09/22/2009,,119762,4928,MARK C-THRU-U BEAUTIFYING SHEER TINT-ALL SHADES,,,4,New Avon LLC,MARK,44,Makeup Products (non-permanent),50,Foundations and Bases,656,13463-67-7,8639,Titanium dioxide,10/14/2009,09-04-13,05/15/2009,10/14/2009,10/14/2009,,119773,4939,MARK LIP GLOSS TRIANGLES-ALL SHADES,,,4,New Avon LLC,MARK,44,Makeup Products (non-permanent),52,Lip Gloss/Shine,656,13463-67-7,8654,Titanium dioxide,10/14/2009,09-04-13,05/15/2010,10/14/2009,10/14/2009,,119925,4998,MARK I-SHEER CREAMY EYE SHADOW HOOK UP-ALL SHADES ,,,4,New Avon LLC,MARK,44,Makeup Products (non-permanent),48,Eye Shadow,656,13463-67-7,8760,Titanium dioxide,10/15/2009,09-05-13,05/15/2010,10/15/2009,10/15/2009,,121609,5790,MARK JUICE GEMS LIP GLOSS (SOLD IN KIT 'JUIC GEMS MINI GIFT SET),,,4,New Avon LLC,AVON,44,Makeup Products (non-permanent),53,"Lip Color - Lipsticks, Liners, and Pencils",656,13463-67-7,9693,Titanium dioxide,10/16/2009,10-02-13,05/15/2010,10/16/2009,10/16/2009,,1

25252,7073,Rouge glossy lipstick,,,301,Yves Rocher Inc.,Luminelle,44,Makeup Products (non-permanent),53,"Lip Color - Lipsticks, Liners, and Pencils",656,13463-67-7,11085,Titanium dioxide,11/19/2009,11-08-13,05/14/2013,11/19/2009,11/19/2009,11/19/2009,125253,7073,Rouge glossy lipstick,,,301,Yves Rocher Inc.,Luminelle,44,Makeup Products (non-permanent),53,"Lip Color - Lipsticks, Liners, and Pencils",656,13463-67-7,11088,Titanium dioxide,11/19/2009,11-08-13,05/14/2013,11/19/2009,11/19/2009,,1

7.7.2 Output data

PrimaryCategory

Makeup Products (non-permanent)

8 Analysis 4

8.1 Analysis Performed

Finding latest 5 removed chemicals in the cosmetics products – In this MapReduce problem, the output will give us the chemicals that are removed to manufacture the cosmetic products.

8.2 Inputs & Output Attributes

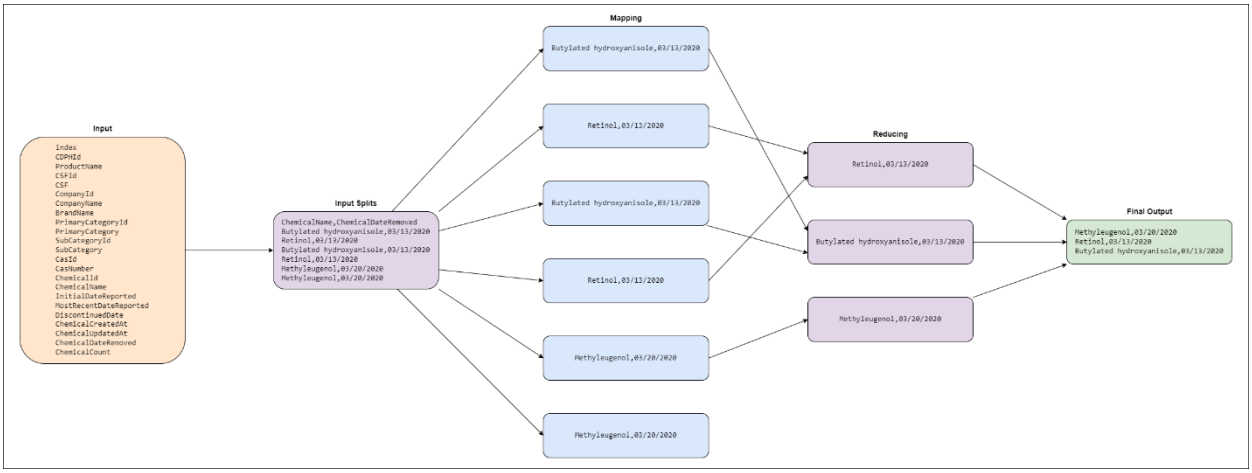
8.2.1 Input Attribute

ChemicalName, *ChemicalDateRemoved*

8.2.2 Outputs Attribute

ChemicalName, *ChemicalDateRemoved*

8.3 MapReduce Diagrams



8.4 Mapper & Reducer Pseudo Codes

8.4.1 Mapper Pseudo Code

```
class Mapper:
        method map(fullColumns):
                columns = fullColumns.split(',')
                if(length(columns) == total_colums)
                        chemicalName = columns[chemicalNamePosition]
                        chemicalDateRemoved = columns[chemicalDateRemovedPosition]
                        write(selected columns)
```

8.4.2 Reducer Pseudo Code

```
class Reducer:
        method reduce(chemicalName, chemicalDateRemoved)
                key = chemicalName
                value = chemicalDateRemoved
                dict = {key:value} # key is string, value is a date field

                if key not in dict:
                        if value > dict.get(key)
                                dict.getValue(key)= value

                        else:
                                dict.getValue(key)= value

                        write(dict.key, dict.value)
```

8.5 Mapper & Reducer Programs

8.5.1 Mapper Program

```
#!/usr/bin/env python3
import sys

delimiter = ","


def map():
   for line in sys.stdin:
      rows = line.strip()
      columns = rows.split(delimiter)

      if len(columns) == 23:
         chemical_name = columns[15]
         chemical_date_removed = columns[21]

         print(f"{chemical_name}{delimiter}{chemical_date_removed}")


if __name__ == "__main__":
   map()
```

8.5.2 Reducer Program

```
#!/usr/bin/env python3
import sys
from datetime import datetime

delimiter = ","
chemicals_removed = {}


def reduce():
    my_iterator = iter(sys.stdin.readline, "")
    header = next(my_iterator)
    chemical_name_header, chemical_date_removed_header = header.strip().split(delimiter)
    print(f"{chemical_name_header}{delimiter}{chemical_date_removed_header}")

    for line in sys.stdin:
        line = line.strip()

        chemical_name, chemical_date_removed = line.split(delimiter)

        if chemical_date_removed is not None and "/" in chemical_date_removed:
            chemical_date_removed = datetime.strptime(chemical_date_removed,
"%m/%d/%Y").strftime("%Y/%m/%d")

            if chemical_name in chemicals_removed.keys():
                if chemical_date_removed > chemicals_removed[chemical_name]:
                    chemicals_removed[chemical_name] = chemical_date_removed
            else:
                chemicals_removed[chemical_name] = chemical_date_removed

    sorted_chemicals_removed = sorted(chemicals_removed.items(), key=lambda kv: (kv[1], kv[0]),
reverse=True)[0:5]

    for chemicals in sorted_chemicals_removed:
        print(f"{chemicals[0]}{delimiter}{chemicals[1]}")


if __name__ == "__main__":
    reduce()
```

8.6 Statistics

```
2022-10-31 07:08:37,526 INFO mapred.LocalJobRunner: Records R/W=1105/1
2022-10-31 07:08:37,526 INFO mapred.Task: Task 'attempt_local362814966_0001_m_000000_0' done.
2022-10-31 07:08:37,532 INFO mapred.Task: Final Counters for attempt_local362814966_0001_m_000000_0: Counters: 17
        File System Counters
                FILE: Number of bytes read=56044453
                FILE: Number of bytes written=30271746
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=114299
                Map output records=63743
                Map output bytes=1191136
                Map output materialized bytes=1318628
                Input split bytes=96
                Combine input records=0
                Spilled Records=63743
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=203
                Total committed heap usage (bytes)=472383488
        File Input Format Counters
                Bytes Read=27950723
2022-10-31 07:08:37,534 INFO mapred.LocalJobRunner: Finishing task: attempt_local362814966_0001_m_000000_0
2022-10-31 07:08:37,534 INFO mapred.LocalJobRunner: map task executor complete.
2022-10-31 07:08:37,537 INFO mapred.LocalJobRunner: Waiting for reduce tasks
```

```
2022-10-31 07:08:37,957 INFO mapred.Task: Task 'attempt_local362814966_0001_r_000000_0' done.
2022-10-31 07:08:37,957 INFO mapred.Task: Final Counters for attempt_local362814966_0001_r_000000_0: Counters: 24
        File System Counters
                FILE: Number of bytes read=58681741
                FILE: Number of bytes written=31590495
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=422
                Reduce shuffle bytes=1318628
                Reduce input records=63743
                Reduce output records=6
                Spilled Records=63743
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=472383488
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=121
2022-10-31 07:08:37,958 INFO mapred.LocalJobRunner: Finishing task: attempt_local362814966_0001_r_000000_0
2022-10-31 07:08:37,958 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-10-31 07:08:38,292 INFO mapreduce.Job:  map 100% reduce 100%
```

```
2022-10-31 07:08:38,293 INFO mapreduce.Job: Job job_local362814966_0001 completed successfully
2022-10-31 07:08:38,302 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=114726194
                FILE: Number of bytes written=61862241
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=114299
                Map output records=63743
                Map output bytes=1191136
                Map output materialized bytes=1318628
                Input split bytes=96
                Combine input records=0
                Combine output records=0
                Reduce input groups=422
                Reduce shuffle bytes=1318628
                Reduce input records=63743
                Reduce output records=6
                Spilled Records=127486
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=203
                Total committed heap usage (bytes)=944766976
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=27950723
        File Output Format Counters
                Bytes Written=121
2022-10-31 07:08:38,304 INFO streaming.StreamJob: Output directory: output
```

8.7 Sample Input & Output data's

8.7.1 Input data

index,CDPHId,ProductName,CSFId,CSF,CompanyId,CompanyName,BrandName,PrimaryCategoryId,PrimaryCategory,SubCategoryId,SubCategory,CasId,CasNumber,ChemicalId,ChemicalName,InitialDateReported,MostRecentDateReported,DiscontinuedDate,ChemicalCreatedAt,ChemicalUpdatedAt,ChemicalDateRemoved,ChemicalCount
103327,37147,A-Zyme Peel,,,1316,Ultraceuticals Pty Ltd,Ultraceuticals,90,Skin Care Products ,105,Other Skin Care Product,92,25013-16-5,61078,Butylated hydroxyanisole,04/23/2019,03-12-20,,04/23/2019,03-12-20,03/13/2020,2
103328,37147,A-Zyme Peel,,,1316,Ultraceuticals Pty Ltd,Ultraceuticals,90,Skin Care Products ,105,Other Skin Care Product,958,68-26-8,67601,Retinol,04/23/2019,03-12-20,,03-12-20,03-12-20,03/13/2020,2
103329,37147,A-Zyme Peel,,,1316,Ultraceuticals Pty Ltd,Ultraceuticals,90,Skin Care Products ,105,Other Skin Care Product,92,25013-16-5,67602,Butylated hydroxyanisole,04/23/2019,03-12-20,,03-12-20,03-12-20,03/13/2020,2
113550,41264,Ultra A Skin Perfecting Serum Mild,,,1316,Ultraceuticals Pty Ltd,Ultraceuticals,90,Skin Care Products ,92,Anti-Wrinkle/Anti-Aging Products (making a cosmetic claim),958,68-26-8,67604,Retinol,03-12-20,03-12-20,,03-12-20,03-12-20,03/13/2020,2
113774,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products ,93,Skin Cleansers,442,93-15-2,67681,Methyleugenol,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,03/20/2020,6
113776,41308,Self Heating Face Mask With Natural Activated Charcoal,64642,Fragrance,1388,Cosmopharm Ltd.,Careline,90,Skin Care Products ,93,Skin Cleansers,442,93-15-2,67683,Methyleugenol,03/20/2020,03/20/2020,,03/20/2020,03/20/2020,03/20/2020,6

8.7.2 Output data

ChemicalName,ChemicalDateRemoved
Methyleugenol,2020/03/20
Retinol,2020/03/13
Butylated hydroxyanisole,2020/03/13

9 References

- Chemicals in Cosmetics
- Running Hadoop on Ubuntu Linux (Single-Node Cluster)
- Setting up a Single Node Cluster
- Writing a Hadoop MapReduce Program in Python
- Source code