Nancy Vien (1895337)
Alejandra Jimenez Nieto (1895997)

Architecture matérielle et logicielle Big Data
420-BD6-BB Gr. 425

Projet – Étude de cas – Printemps 2018

Travail présenté
à Nesrine Zemirli

Collège de Bois-de-Boulogne
26 jul 2018

1. Introduction

Notre application consiste à développer un système de prévision de l'asthme et des allergies respiratoires au Canada. Elle permettra de faire une prévision des allergies respiratoires ou encore l'asthme à une localisation précise et une période spécifique. L'application peut aussi prévenir ou éviter les symptômes d'une personne déjà allergique en un endroit spécifique et à un moment donné et ainsi empêcher que cette allergie ne se complique en asthme ou encore en une autre maladie respiratoire.

Exemples de cas d'utilisation :

Utilisateur non-inscrit

- Donne des informations pour prévenir les allergies respiratoires et asthme pour la population en général.
- Donne des recommandations pour éviter que des allergies se développent.

Utilisateur non allergique inscrit

- L'utilisateur répond à une liste de question pour étoffer son dossier.
- Système de recommandations personnalisées géolocalisées pour les abonnées allergiques.
- Le résultat obtenu aide à formuler différentes solutions pour prévenir quelques situations reliées à l'allergie.

Utilisateur ayant déjà une allergie connue ou l'asthme

- L'utilisateur entre des données personnelles journalières pour l'aider à maitriser et contrôler ses symptômes
- L'utilisateur reçoit des alertes personnalisées pour l'aider au contrôle et la planification de son allergie
- L'utilisateur entre des informations précises (questionnaire étoffé) et sur une base historique et contextuel l'application pourra lui prédire s'il développera une allergie ou encore une maladie respiratoire ou encore son pourcentage de risque.
- Les résultats obtenus donneront des recommandations pour éviter que l'allergie connue se complique en maladie respiratoires comme l'asthme ou la rhinite chronique
- Permet de prévenir l'asthme ou la rhinite chronique
- Permet d'éviter une crise d'asthme

2. Données

2.1 Données

Nous disposons d'un ensemble de données provenant de plusieurs sources. Ces données sont semi-structurées en format CSV. Elles portent sur plusieurs aspects dont plusieurs facteurs climatiques,

en voici quelques-uns :

2.1.1 Indicateurs de la qualité de l'air[1].

Les deux principaux polluants qui composent le smog sont l'ozone troposphérique et les particules (PT : Particules Totales d'un diamètre maximal d'environ 100 micromètres, P10 : Particules d'un diamètre inférieur à 10 micromètres, P2.5 : Particules d'un diamètre inférieures à 2.5 micromètres). Les particules sont des fragments en suspension dans l'air sous forme solide ou liquide[2]. La taille des particules détermine la portée des dommages que celles-ci causent à la sante (il y a un lien entre les particules et de troubles respiratoires tels que l'asthme, la bronchite et l'emphysème).

La forte densité de smog est associée à la saison estivale pour la présence de rayons du soleil et des températures élevées. Le smog hivernal (à cause de la contribution des particules au lieu de l'ozone) étant une préoccupation sérieuse, quand l'air en stagnation accumule les polluants provenant du chauffage au bois et de l'utilisation de véhicules[3].

Les indicateurs de la qualité de l'air suivent les mesures de l'exposition des Canadiens à l'ozone troposphérique et aux particules fines (PM2,5). Les indicateurs d'exposition à l'ozone et aux P2,5 sont des concentrations moyennes pondérées en fonction de la population observées dans les stations de surveillance au Canada pendant la saison chaude (d'avril à septembre)[4].

2.1.2 Inventaire d'émissions polluantes atmosphériques[5] :

Monoxyde de carbone (CO), Composés organiques volatils (COV).

2.1.3 Asthme en Canada

Décès selon la cause : Maladies de l'appareil respiratoire[6] , Personnes avec maladie pulmonaire obstructive chronique[7].

2.1.4 Données climatiques

En tenant compte du fait qu'il existe une relation directe entre les maladies respiratoires et le climat, nous utiliserons des données historiques climatiques. Il y a différentes sources, par exemple sources qui viennent de différentes stations de mesure[8] :

---

[1] https://www.canada.ca/fr/environnement-changement-climatique/services/indicateurs-environnementaux/qualite-air.html
[2] https://www.canada.ca/fr/environnement-changement-climatique/services/pollution-atmospherique/polluants/principaux-contaminants/matieres-particulaires.html
[3] https://www.canada.ca/fr/environnement-changement-climatique/services/pollution-atmospherique/enjeux/smog-causes-effets.html
[4] http://www.statcan.gc.ca/pub/16-251-x/16-251-x2007000-eng.htm
[5] https://www.canada.ca/fr/environnement-changement-climatique/services/pollution-atmospherique/publications/inventaire-emissions-polluants-atmospheriques-2016/chapitre-2-5.html
[6] http://www5.statcan.gc.ca/cansim/a26?lang=fra&id=1020530&p2=33#F2
[7] http://www.statcan.gc.ca/tables-tableaux/sum-som/l02/cst01/health104a-fra.htm

3. Développement

Dans cette première phase de notre application, nous avons déterminé quelques questions auxquelles nous aimerions répondre.

Pour les fins de notre exemple, le traitement des données se fera à partir d'un job MapReduce avec le framework MrJob pour Hadoop.

3.1 Collecte, transformation et chargement l'information

3.1.1 Première question concernant les taux d'incidence d'asthme

Dataset utilisé :
Taux d'incidence d'Asthme en Canada[1].

Question :
Quelle province canadienne a eu le plus haut taux d'incidence asthmatique entre 2000 et 2012 ?

> Un incident d'asthme est identifié au cours de l'année où la personne répondait à la définition de cas pour la première fois[2]. Le taux est calculé avec le nombre de cas (incidents) d'asthme sur le total de la population.

Après l'exécution du code Python avec le libraire MRJob (voir Annexe 1), nous avons obtenu le résultat suivant :

> *Streaming final output from /tmp/maximum_annee.cloudera.20180725.035437.123117/output...*
> ***21.09 ["Prince Edward Island (PE)", "Males", "2008"]***

Réponse :
Nous pouvons constater que le plus haut taux d'incidence a été enregistré à l'Île-du-Prince-Édouard en 2008 chez les hommes.

3.1.2. Deuxième question concernant les taux d'incidence d'asthme

Dataset utilisé :
Taux d'incidence d'Asthme en Canada[3].

[1] https://infobase.phac-aspc.gc.ca/ccdss-scsmc/data-tool/
[2] https://www.canada.ca/fr/sante-publique/services/publications/maladies-et-affections/asthme-maladie-pulmonaire-obstructive-chronique-canada-2018.html
[3] https://infobase.phac-aspc.gc.ca/ccdss-scsmc/data-tool/

Question :
Dans quelle province (et en quelle année) les taux d'incidence de l'asthme sont-ils supérieurs à 20% entre 2000 et 2012 ?

Après l'exécution du code Python avec le libraire MRJob (voir Annexe 2), nous avons obtenu le résultat suivant :

```
!python projet/maximum_annee.py /home/cloudera/projet/PHAC_Infobase_CCDSS.csv
No configs found; falling back on auto-configuration
Creating temp directory /tmp/maximum_annee.cloudera.20180725.202856.486539
Running step 1 of 1...
Streaming final output from
/tmp/maximum_annee.cloudera.20180725.202856.486539/output...
[20.17, "2007", "Males", "Ontario"]     20.17
[20.2, "2010", "Males", "Nova Scotia (NS)"]     20.2
[20.34, "2006", "Males", "Nova Scotia (NS)"]    20.34
[20.35, "2012", "Males", "Ontario"]     20.35
[20.4, "2008", "Males", "Ontario"]      20.4
[20.43, "2005", "Males", "Prince Edward Island (PE)"]   20.43
[20.57, "2007", "Males", "Nova Scotia (NS)"]    20.57
[20.57, "2011", "Males", "Ontario"]     20.57
[20.6, "2009", "Males", "Nova Scotia (NS)"]     20.6
[20.6, "2009", "Males", "Ontario"]      20.6
[20.61, "2011", "Males", "Prince Edward Island (PE)"]   20.61
[20.67, "2010", "Males", "Ontario"]     20.67
[20.7, "2008", "Males", "Nova Scotia (NS)"]     20.7
[20.73, "2012", "Males", "Prince Edward Island (PE)"]   20.73
[20.86, "2009", "Males", "Prince Edward Island (PE)"]   20.86
[21.0, "2006", "Males", "Prince Edward Island (PE)"]    21.0
[21.0, "2010", "Males", "Prince Edward Island (PE)"]    21.0
[21.01, "2007", "Males", "Prince Edward Island (PE)"]   21.01
[21.09, "2008", "Males", "Prince Edward Island (PE)"]   21.09
```

Réponse :
Les provinces ayant eu un taux d'incidence supérieur à 20% (par genre) sont majoritairement l'Île-du-Prince-Édouard, suivi de l'Ontario et de la Nouvelle Écosse entre 2000 à 2012.

3.1.3 Troisième question concernant la maladie pulmonaire obstructive chronique

Dataset utilisé :

Taux d'incidence de la maladie pulmonaire obstructive chronique (MPOC) en Canada[4].

Un incident de MPOC est identifié au cours de l'année où la personne répondait à la définition de cas pour la première fois[5]. Le taux est calculé avec le nombre de cas (incidents) d'asthme sur le total de la population.

Question :
Quelle province canadienne a eu le plus taux d'incidence de MPOC entre 2000 et 2012 ?

Après l'exécution du code Python avec le libraire MRJob (voir Annexe 1), on a obtenu le résultat suivant :

!python projet/maximum_t.py /home/cloudera/projet/PHAC_Infobase_CCDSS_MPOC.csv
No configs found; falling back on auto-configuration
Creating temp directory /tmp/maximum_t.cloudera.20180725.215008.388591
Running step 1 of 2...
Running step 2 of 2...
Streaming final output from /tmp/maximum_t.cloudera.20180725.215008.388591/output...
57.04 ["Nunavut (NU)", "Females", "2005"]
Removing temp directory /tmp/maximum_t.cloudera.20180725.215008.388591..

Réponse :
Le plus grand taux d'incidence a été enregistré (par genre) au Nunavut en 2005 chez les femmes.

3.1.4 Quatrième question concernant la maladie pulmonaire obstructive chronique

Dataset utilisé :
Taux d'incidence de la maladie pulmonaire obstructive chronique (MPOC) en Canada[6].

Question :
Dans quelle province (et en quelle année) les taux d'incidence de MPOC sont-ils supérieurs à 20% entre 2000 et 2012 ?

Après l'exécution du code Python avec le libraire MRJob (voir Annexe 2), on a obtenu le résultat suivant :

!python projet/maximum_annee.py /home/cloudera/projet/PHAC_Infobase_CCDSS_MPOC.csv
No configs found; falling back on auto-configuration

[4] https://infobase.phac-aspc.gc.ca/ccdss-scsmc/data-tool/
[5] https://www.canada.ca/fr/sante-publique/services/publications/maladies-et-affections/asthme-maladie-pulmonaire-obstructive-chronique-canada-2018.html#a2.2
[6] https://infobase.phac-aspc.gc.ca/ccdss-scsmc/data-tool/

Creating temp directory /tmp/maximum_annee.cloudera.20180725.214824.632373
Running step 1 of 1...
Streaming final output from
/tmp/maximum_annee.cloudera.20180725.214824.632373/output...
[20.19, "2009", "Males", "Nunavut (NU)"] 20.19
[20.52, "2011", "Males", "Nunavut (NU)"] 20.52
[21.02, "2011", "Both sexes", "Nunavut (NU)"] 21.02
[21.9, "2011", "Females", "Nunavut (NU)"] 21.9
[22.22, "2000", "Females", "Northwest Territories (NT)"] 22.22
[23.35, "2012", "Both sexes", "Nunavut (NU)"] 23.35
[25.75, "2007", "Males", "Nunavut (NU)"] 25.75
[25.79, "2010", "Both sexes", "Nunavut (NU)"] 25.79
[27.84, "2007", "Both sexes", "Nunavut (NU)"] 27.84
[28.4, "2012", "Females", "Nunavut (NU)"] 28.4
[30.0, "2009", "Both sexes", "Nunavut (NU)"] 30.0
[31.02, "2007", "Females", "Nunavut (NU)"] 31.02
[31.16, "2008", "Males", "Nunavut (NU)"] 31.16
[32.05, "2008", "Both sexes", "Nunavut (NU)"] 32.05
[32.4, "2008", "Females", "Nunavut (NU)"] 32.4
[33.33, "2010", "Females", "Nunavut (NU)"] 33.33
[40.29, "2009", "Females", "Nunavut (NU)"] 40.29
[44.31, "2006", "Females", "Nunavut (NU)"] 44.31
[47.12, "2006", "Both sexes", "Nunavut (NU)"] 47.12
[47.21, "2005", "Males", "Nunavut (NU)"] 47.21
[50.41, "2006", "Males", "Nunavut (NU)"] 50.41
[51.55, "2005", "Both sexes", "Nunavut (NU)"] 51.55
[57.04, "2005", "Females", "Nunavut (NU)"] 57.04
Removing temp directory /tmp/maximum_annee.cloudera.20180725.214824.632373...

Réponse :
La province la plus affectée par les incidents de MPOC (par genre et sans genre) est le Nuvavut, entre 2000 et 2012 chez les hommes et les femmes.

3.1.5 Cinquième question concernant les polluants atmosphériques

Dataset utilisé :
Inventaire national des rejets de polluants (INRP) - Données normalisées

L'INRP est l'inventaire public du Canada sur les polluants rejetés dans l'atmosphère, dans l'eau et dans le sol, éliminés et transférés aux fins de recyclage[7]. Nous avons pris la source de données «NPRI-SubsRele-Normalized-Since1994.csv» qui contient toutes les quantités de rejet de substances.

[7] https://ouvert.canada.ca/data/fr/dataset/40e01423-7728-429c-ac9d-2954385ccdfb?_ga=2.197817644.2017876941.1532560051-1526959871.1532560051

Nous avons limité notre questionnement aux polluants rejetés dans l'air, qui sont les plus importants composants du smog, soit P10, P2.5 et le monoxyde de carbone.

Question :
Quelle province canadienne a rejeté la plus grande quantité de polluant (soit P10, P2.5 et monoxyde de carbone) dans l'air entre 2000 et 2012 ?

Après l'exécution du code Python avec le libraire MRJob (voir Annexe 3), on a obtenu le résultat suivant :

438248.29750000004 ["2011", "QC", "Carbon monoxide", "tonnes"]

Réponse :
La province ayant rejetée la plus grande quantité de polluant dans l'air, soit le monoxyde de carbone, 438248.29750000004 tonnes est le Québec en 2011.

3.1.6 Sixième question concernant les polluants atmosphériques

Dataset utilisé :
Inventaire national des rejets de polluants (INRP) - Données normalisées

Question :
Quelle sont les quantités rejetées de polluants (P10, P2.5 et monoxyde de carbone), par province au Canada entre 2010 et 2012 ?

Réponse :
La réponse à cette question est plutôt longue, il faut donc se référer à l'Annexe 4 pour connaitre la quantité de polluant de P10, P2.5 et de monoxyde de carbone par province. Nous pouvons quand même nommer ici les provinces qui se retrouve en tête de liste pour chaque polluant.

La province qui rejette le plus de P10 est la Colombie britannique.
La province qui rejette le plus de P2.5 est l'Alberta suivi de près par la Colombie britannique.
La province qui rejette le plus de monoxyde de carbone reste le Québec comme stipulé à la question précédente.

3.1.7 Septième question concernant les polluants atmosphériques

Dataset utilisé :
Inventaire national des rejets de polluants (INRP) - Données normalisées

Question :
Quelle sont les moyennes de polluant d'ozone, par province en Canada entre les années 2010 et 2012 ?

Le code python pour répondre cette question se trouve dans l'Annexe 5.
Le résultat est le suivant :

!python projet/pollution_o3.py /home/cloudera/projet/2012_AnnualO3_v5.csv
No configs found; falling back on auto-configuration
Creating temp directory /tmp/pollution_o3.cloudera.20180726.185453.790982
Running step 1 of 1...
Streaming final output from /tmp/pollution_o3.cloudera.20180726.185453.790982/output...
["AB", "2010"] 617.0
["AB", "2011"] 756.0
["AB", "2012"] 800.0
["BC", "2010"] 632.0
["BC", "2011"] 694.0
["BC", "2012"] 759.0
["MB", "2010"] 74.0
["MB", "2011"] 77.0
["MB", "2012"] 76.0
["NB", "2010"] 318.0
["NB", "2011"] 313.0
["NB", "2012"] 304.0
["NL", "2010"] 130.0
["NL", "2011"] 174.0
["NL", "2012"] 209.0
["NS", "2010"] 111.0
["NS", "2011"] 251.0
["NS", "2012"] 190.0
["NT", "2010"] 110.0
["NT", "2011"] 152.0
["NT", "2012"] 108.0
["ON", "2010"] 1355.0
["ON", "2011"] 1291.0
["ON", "2012"] 1325.0
["QC", "2010"] 1239.0
["QC", "2011"] 1190.0
["QC", "2012"] 1209.0
["SK", "2010"] 116.0
["SK", "2011"] 135.0
["SK", "2012"] 121.0
Removing temp directory /tmp/pollution_o3.cloudera.20180726.185453.790982..

Réponse :

Les émissions d'ozone les plus élevées, entre 2010 à 2012 , se retrouvent dans les provinces de l'Ontario (en moyenne 1323 ppb) et du (en moyenne 1212 ppb).

4. Conclusion

Ce premier exercice nous a permis de répondre à certaines questions. Il semble trop tôt à cette étape d'en tirer des conclusions définitives mais nous pouvons nous avancer sur certaines pistes ou hypothèses.

Les données concernant l'asthme nous permettent de dire que les hommes sont plus touchés par l'asthme que les femmes. Et, les provinces de l'est ont les taux les plus élevés d'incidence asthmatique, des taux supérieurs à 20%.

Pour le MPOC, bien que le taux d'incidence fluctue d'une année à l'autre, le Nunavut est la seule province, entre 2000 et 2012 à avoir des taux supérieurs à 20% pour atteindre des taux jusqu'à 57% certaines années. Nous ne pouvons conclure que les femmes sont plus touchées par le MPOC puisque cela change d'année en année où parfois ce sont les hommes qui ont un taux d'incidence supérieurs aux femmes et d'autres années c'est le contraire. On peut par contre voir la tendance d'un taux qui se maintient autour de 20% à 25% dans les dernières années soit 2010, 2011 et 2012. Ce serait intéressant d'avoir les données pour les années de 2013 à 2017 pour vérifier si la tendance se maintient.

Pour certains polluants atmosphériques, ce ne sont pas les provinces les plus peuplées qui sont nécessairement en tête de liste. Il se peut que pour certains d'entre eux, leur quantité soit proportionnel à la population mais ce n'est pas le cas pour certains comme pour le P10 et le p2.5 où les quantités émissent les plus élevés sont en Colombie britannique et en Alberta où la population correspond à plus ou moins la moitié ou moins de la moitié des grandes provinces comme l'Ontario ou le Québec.
Il serait intéressant de voir le ratio population pour vérifier si certaines quantités de polluants sont bel et bien proportionnel à leur population. Ce qui semble le cas pour les polluants d'ozone.

Annexe 1

```
'''Obtenir le taux d'asthme plus eleve par annee dans les provinces du Canada'''
from mrjob.job import MRJob
from mrjob.step import MRStep

class MRFriendsByAge(MRJob):

    def steps(self):
        return [
            MRStep(mapper=self.mapper,
                   reducer=self.reducer),
                    MRStep(mapper=self.mapper_2,
                     reducer=self.reducer_2)
        ]
    def mapper(self, _, line):
        fields = line.split(',')
        place=fields[0]
        Age = fields[2]
        Sex = fields[3]
        year = fields[4]
        rate = fields[5]
        yield (place,Sex,year), float(rate)

    def reducer(self, key, rate):

        yield key, max(rate)

    def mapper_2(self, place, maxRate):
        yield None, (maxRate, place)

    def reducer_2(self, key, value):
        yield max(value)

if __name__ == '__main__':
    MRFriendsByAge.run()
```

Annexe 2

```
# -*- coding: utf-8 -*-
'''Obtenir le taux d'asthme plus élevé par année par province en Canada'''
from mrjob.job import MRJob
from mrjob.step import MRStep

class MRFriendsByAge(MRJob):
    SORT_VALUES = True
    def steps(self):
        return [
            MRStep(mapper=self.mapper,
            reducer=self.reducer)
        ]
    def mapper(self, _, line):
        fields = line.split(',')
        if (float(fields[5])>20):
            yield (float(fields[5]),fields[4],fields[3],fields[0]), float(fields[5])

    def reducer(self, key, rate):
        yield key, max(rate)


if __name__ == '__main__':
    MRFriendsByAge.run()
```

Annexe 3

```
# -*- coding: utf-8 -*-

from mrjob.job import MRJob
from mrjob.step import MRStep

class MRPollution(MRJob):
   SORT_VALUES = True
   def steps(self):
      return [
         MRStep(mapper=self.mapper,
         reducer=self.reducer),
         MRStep(mapper=self.mapper_2,
                reducer=self.reducer_2)
      ]
   def mapper(self, _, line):
      fields = line.split(',')
      if ((unicode(fields[3], errors='ignore')=="Total Air" or
         unicode(fields[3], errors='ignore')=="Total All Media<1t") and
         unicode(fields[4], errors='ignore')!="" and
         (unicode(fields[2], errors='ignore')=="Carbon monoxide" or
         unicode(fields[2], errors='ignore')=="PM10" or
         unicode(fields[2], errors='ignore')=="PM2.5")):
         yield (unicode(fields[0], errors='ignore'),unicode(fields[1],
errors='ignore'),unicode(fields[2], errors='ignore'),unicode(fields[5],
errors='ignore')),float(fields[4])

   def reducer(self, key, rate):
      yield key, sum(rate)

   def mapper_2(self, key, maxRate):
      yield None, (maxRate, key)

   def reducer_2(self, key, value):
      yield max(value)


if __name__ == '__main__':
   MRPollution.run()
```

Annexe 4

Programme :

```
# -*- coding: utf-8 -*-

from mrjob.job import MRJob
from mrjob.step import MRStep

class MRPollution(MRJob):
   SORT_VALUES = True
   def steps(self):
      return [
         MRStep(mapper=self.mapper,
         reducer=self.reducer)
      ]
   def mapper(self, _, line):
      fields = line.split(',')
      if ((unicode(fields[3], errors='ignore')=="Total Air" or
         unicode(fields[3], errors='ignore')=="Total All Media<1t") and
         unicode(fields[4], errors='ignore')!="" and
         (unicode(fields[2], errors='ignore')=="Carbon monoxide" or
         unicode(fields[2], errors='ignore')=="PM10" or
         unicode(fields[2], errors='ignore')=="PM2.5")
         and unicode(fields[0], errors='ignore')>"2009" and
         unicode(fields[0], errors='ignore')<"2013"):
         yield (unicode(fields[2], errors='ignore'),unicode(fields[0],
errors='ignore'),unicode(fields[1], errors='ignore'),unicode(fields[5],
errors='ignore')),float(fields[4])

   def reducer(self, key, rate):
      yield key, sum(rate)



if __name__ == '__main__':
   MRPollution.run()
```

Reponse :

```
!python projet/pollution.py /home/cloudera/projet/NPRI-SubsReleA.csv
No configs found; falling back on auto-configuration
Creating temp directory /tmp/pollution.cloudera.20180726.194223.906225
Running step 1 of 1...
Streaming final output from /tmp/pollution.cloudera.20180726.194223.906225/output...
["Carbon monoxide", "2010", "AB", "tonnes"]     229557.8642999999
["Carbon monoxide", "2010", "BC", "tonnes"]     126208.86099999993
["Carbon monoxide", "2010", "MB", "tonnes"]     1757.0561999999998
```

["Carbon monoxide", "2010", "NB", "tonnes"] 25686.414300000004
["Carbon monoxide", "2010", "NL", "tonnes"] 9812.600400000001
["Carbon monoxide", "2010", "NS", "tonnes"] 3898.3103999999994
["Carbon monoxide", "2010", "NT", "tonnes"] 1741.4485
["Carbon monoxide", "2010", "NU", "tonnes"] 875.9540000000002
["Carbon monoxide", "2010", "ON", "tonnes"] 78133.68439999998
["Carbon monoxide", "2010", "PE", "tonnes"] 44.473
["Carbon monoxide", "2010", "QC", "tonnes"] 416144.40189999976
["Carbon monoxide", "2010", "SK", "tonnes"] 21818.719999999994
["Carbon monoxide", "2011", "AB", "tonnes"] 197983.7511000001
["Carbon monoxide", "2011", "BC", "tonnes"] 123358.25589999993
["Carbon monoxide", "2011", "MB", "tonnes"] 1844.879
["Carbon monoxide", "2011", "NB", "tonnes"] 16689.197000000004
["Carbon monoxide", "2011", "NL", "tonnes"] 7703.5561
["Carbon monoxide", "2011", "NS", "tonnes"] 4460.013999999999
["Carbon monoxide", "2011", "NT", "tonnes"] 2267.188
["Carbon monoxide", "2011", "NU", "tonnes"] 832.16
["Carbon monoxide", "2011", "ON", "tonnes"] 72484.43700000002
["Carbon monoxide", "2011", "PE", "tonnes"] 47.807
["Carbon monoxide", "2011", "QC", "tonnes"] 438248.29750000004
["Carbon monoxide", "2011", "SK", "tonnes"] 21362.608300000007
["Carbon monoxide", "2011", "YT", "tonnes"] 61.9231
["Carbon monoxide", "2012", "AB", "tonnes"] 231253.7800000001
["Carbon monoxide", "2012", "BC", "tonnes"] 131327.88620000007
["Carbon monoxide", "2012", "MB", "tonnes"] 2218.3450000000003
["Carbon monoxide", "2012", "NB", "tonnes"] 17951.7019
["Carbon monoxide", "2012", "NL", "tonnes"] 9091.248200000002
["Carbon monoxide", "2012", "NS", "tonnes"] 3501.0366999999997
["Carbon monoxide", "2012", "NT", "tonnes"] 2266.6204000000002
["Carbon monoxide", "2012", "NU", "tonnes"] 634.0896000000001
["Carbon monoxide", "2012", "ON", "tonnes"] 74711.27479999996
["Carbon monoxide", "2012", "PE", "tonnes"] 84.79
["Carbon monoxide", "2012", "QC", "tonnes"] 413963.8646000001
["Carbon monoxide", "2012", "SK", "tonnes"] 14976.802499999998
["PM10", "2010", "AB", "tonnes"] 34763.5949
["PM10", "2010", "BC", "tonnes"] 48597.867799999985
["PM10", "2010", "MB", "tonnes"] 2386.0440000000012
["PM10", "2010", "NB", "tonnes"] 2538.6924999999997
["PM10", "2010", "NL", "tonnes"] 4692.406
["PM10", "2010", "NS", "tonnes"] 3761.778099999999
["PM10", "2010", "NT", "tonnes"] 1200.9673
["PM10", "2010", "NU", "tonnes"] 609.035
["PM10", "2010", "ON", "tonnes"] 17393.107799999987
["PM10", "2010", "PE", "tonnes"] 68.34
["PM10", "2010", "QC", "tonnes"] 18046.1665
["PM10", "2010", "SK", "tonnes"] 8154.929600000001
["PM10", "2010", "YT", "tonnes"] 4.5122
["PM10", "2011", "AB", "tonnes"] 34587.94860000001
["PM10", "2011", "BC", "tonnes"] 60321.482400000015
["PM10", "2011", "MB", "tonnes"] 2290.333800000001
["PM10", "2011", "NB", "tonnes"] 2500.8381000000004
["PM10", "2011", "NL", "tonnes"] 4022.5427999999993
["PM10", "2011", "NS", "tonnes"] 3316.626300000001
["PM10", "2011", "NT", "tonnes"] 1486.2939999999999
["PM10", "2011", "NU", "tonnes"] 575.226

["PM10", "2011", "ON", "tonnes"] 18333.522999999997
["PM10", "2011", "PE", "tonnes"] 72.633
["PM10", "2011", "QC", "tonnes"] 17188.983900000003
["PM10", "2011", "SK", "tonnes"] 8509.207000000004
["PM10", "2011", "YT", "tonnes"] 6.819
["PM10", "2012", "AB", "tonnes"] 35110.885700000006
["PM10", "2012", "BC", "tonnes"] 78471.82139999999
["PM10", "2012", "MB", "tonnes"] 2556.0045999999998
["PM10", "2012", "NB", "tonnes"] 2396.0664999999995
["PM10", "2012", "NL", "tonnes"] 4414.9057
["PM10", "2012", "NS", "tonnes"] 3144.4474000000005
["PM10", "2012", "NT", "tonnes"] 797.5158000000001
["PM10", "2012", "NU", "tonnes"] 1006.285
["PM10", "2012", "ON", "tonnes"] 17753.35750000002
["PM10", "2012", "PE", "tonnes"] 44.817
["PM10", "2012", "QC", "tonnes"] 18658.931600000004
["PM10", "2012", "SK", "tonnes"] 7956.296800000002
["PM2.5", "2010", "AB", "tonnes"] 11836.991500000011
["PM2.5", "2010", "BC", "tonnes"] 11570.322500000004
["PM2.5", "2010", "MB", "tonnes"] 1180.1277
["PM2.5", "2010", "NB", "tonnes"] 1051.8684
["PM2.5", "2010", "NL", "tonnes"] 1922.3402000000003
["PM2.5", "2010", "NS", "tonnes"] 1893.5643
["PM2.5", "2010", "NT", "tonnes"] 488.81520000000006
["PM2.5", "2010", "NU", "tonnes"] 315.37899999999996
["PM2.5", "2010", "ON", "tonnes"] 9037.0019
["PM2.5", "2010", "PE", "tonnes"] 41.268
["PM2.5", "2010", "QC", "tonnes"] 10113.764299999993
["PM2.5", "2010", "SK", "tonnes"] 3211.821899999998
["PM2.5", "2010", "YT", "tonnes"] 4.4645
["PM2.5", "2011", "AB", "tonnes"] 10986.360199999992
["PM2.5", "2011", "BC", "tonnes"] 11047.208400000003
["PM2.5", "2011", "MB", "tonnes"] 1385.8477999999998
["PM2.5", "2011", "NB", "tonnes"] 1093.706
["PM2.5", "2011", "NL", "tonnes"] 1596.32
["PM2.5", "2011", "NS", "tonnes"] 1801.1612999999998
["PM2.5", "2011", "NT", "tonnes"] 405.121
["PM2.5", "2011", "NU", "tonnes"] 159.032
["PM2.5", "2011", "ON", "tonnes"] 9759.927900000004
["PM2.5", "2011", "PE", "tonnes"] 44.818000000000005
["PM2.5", "2011", "QC", "tonnes"] 9896.479399999995
["PM2.5", "2011", "SK", "tonnes"] 3477.491100000001
["PM2.5", "2011", "YT", "tonnes"] 7.0445
["PM2.5", "2012", "AB", "tonnes"] 10913.345599999999
["PM2.5", "2012", "BC", "tonnes"] 12813.218
["PM2.5", "2012", "MB", "tonnes"] 1232.3986
["PM2.5", "2012", "NB", "tonnes"] 1064.7844999999998
["PM2.5", "2012", "NL", "tonnes"] 1649.2769999999998
["PM2.5", "2012", "NS", "tonnes"] 1664.3497999999995
["PM2.5", "2012", "NT", "tonnes"] 411.53710000000007
["PM2.5", "2012", "NU", "tonnes"] 153.561
["PM2.5", "2012", "ON", "tonnes"] 9471.9345
["PM2.5", "2012", "PE", "tonnes"] 25.811
["PM2.5", "2012", "QC", "tonnes"] 10166.043999999998
["PM2.5", "2012", "SK", "tonnes"] 3594.0489000000002

["PM2.5", "2012", "YT", "tonnes"]	0.4621
Removing temp directory /tmp/pollution.cloudera.20180726.194223.906225...

Annexe 5

```
from mrjob.job import MRJob
from mrjob.step import MRStep

class MRPollution(MRJob):
   SORT_VALUES = True
   def steps(self):
      return [
         MRStep(mapper=self.mapper,
         reducer=self.reducer)
      ]
   def mapper(self, _, line):
      fields = line.split(',')
      if fields[len(fields)-4]!="":
         yield (unicode(fields[0], errors='ignore'),unicode(fields[len(fields)-1],
errors='ignore')),float(fields[len(fields)-4])

   def reducer(self, key, rate):
      yield key, sum(rate)



if __name__ == '__main__':
   MRPollution.run()
```