1. Récupérer les données

Vous devez récupérez les données à partir de:

http://hadoop-master.plg.site.univ-lorraine.fr:50070/explorer.html#/data

Le nom du fichier de données est `Crimes-2001-present.csv`.

2. Format des données

- ID
- case number
- date
- block
- IUCR
- primary type
- description
- location description
- arrest
- domestic
- beat
- district
- ward
- community area
- FBI code
- x coordinate
- y coordinate
- year
- update on
- latitude
- longitude
- location

3. Travail demandé

Vous devez fournir les différents programmes Map Reduce permettant de répondre aux questions suivantes:

1. Donnez le classement décroissant des catégories de crimes
2. Donnez le nombre de crimes en fonction de 6 plages horaires (0-4, 4-8, 8-12, 12-16, 16-20, 20-24)
3. Donnez les 3 zones les plus dangereuses et les zones les moins dangereuses (rayon de 2 kms)
4. Donnez la répartition géographique des crimes commis/élucidés (arrest)
5. Donnez le top 3 des mois les plus concernés par les cas de crimes

On vous demande de présenter les résultats de façon graphique. Les graphiques devront être clairs et devront permettre de rapidement répondre aux questions ci-dessus.

Vous présenterez également les données techniques (temps de traitement), et les hypothèses que vous aurez été amnenés à faire.

4. Rappels

Avant de se lancer dans le traitement complet du jeu de données, on teste généralement sur un échantillon réduit. Vous devrez écrire un jeu de test qui intègre les différents cas mentionnés ci-dessus.

Il est également conseillé de commencer par le mode local, puis de tester sur le mode pseudo-distribué afin de valider complètement votre architecture, avant de lancer vos programmes sur le cluster.