PSDG : Projet Hadoop

Antoine Haas - Jason Golda

Proposition de sujet

Jeu de données : Open data de la ville de Strasbourg (https://www.strasbourg.eu/open-data)

Plus précisément sera utilisé le jeu de données du trafic routier de la CUS, disponible à l'adresse https://www.strasbourg.eu/trafic-routier-eurometropole.

Le jeu de données étant mis à jour toutes les 3 minutes un programme de récupération des données se chargera de récupérer le jeu de données pendant chaque période et ceci sur une période de plusieurs jours (à priori 1 semaine de travail donc du lundi au vendredi).

Questions initiales :

Trouver le temps de trajet moyen entre un point A et un point B sur 1 semaine Identifier les zones les plus lentes à un instant T Identifier les zones les plus lentes en moyenne sur la semaine ou sur plusieurs jours ou sur un seul jour

Modification du sujet initial

La question initiale traitant du temps de trajet moyen n'a pas pu être faite car trop complexe en pré-traitement des données, du fait que les données récupérées sont à transformer en graphe pour pouvoir être traitées.

Afin de pouvoir pallier à ce manque, 2 questions ont néanmoins été rajoutées qui sont :

- Calcul de la somme des états pour chacun des capteurs
- Calcul de la moyenne des états pour chaque heure - minute

Questions à traitées

On se retrouver donc avec les questions ci-dessous à traiter :

1. Identifier les zones les plus lentes à un instant T
2. Identifier les zones les plus lentes en moyenne sur la semaine ou sur plusieurs jours ou sur un seul jour
3. Calcul de la somme des états pour chacun des capteurs
4. Calcul de la moyenne des états pour chaque heure - minute

Pré-traitement des données

Téléchargement des données

Les données de la CUS étant mises à jour toutes les 3 minutes, un travail de collecte des données a été fait au préalable afin de pouvoir récupérer un maximum de données exploitable sur une plage de 1 semaine. Le script de récupération des données est `collect-dataset.ps1` pour windows et `collect-dataset.sh` pour mac / linux.

Regroupement des données téléchargées

Hadoop n'étant pas adapté au traitement de grand nombre de fichiers et des fichiers XML en eux-même, un script python a été développé afin de parser les différents fichiers XML qui ont été téléchargés et les regrouper en un seul fichier CSV, facilitant ainsi le traitement par la suite avec Hadoop.

Pendant cette étape de parsing, la liaison entre le nom et l'ID du capteur est également faite (nom récupéré depuis le fichier cus-traffic.gml).

Quelques infos sur les données

Une fois regroupé, le CSV contient 4 251 645 lignes, pour une récupération effectuée, sans interruption, entre le 07 mars à 22h12 et le 14 mars à 12h32.

Algorithmes

Zone les plus lentes

```
Pour chaque ligne du CSV
    Grouper les lignes par nom : débit

Pour chaque groupe
    Sommer les débits
```

Zone les plus lentes en moyenne sur une journée

```
Pour chaque ligne du CSV
    Grouper les lignes par date|nom : débit

Pour chaque groupe de date|nom
    Sommer les débits

Pour chaque ligne date|nom
    Grouper par nom

Pour chaque groupe de nom
    Diviser débit total / nombre de date
```

Somme des états par capteur

```
Pour chaque ligne du CSV
    Grouper les lignes par nom : état

Pour chaque groupe de nom
    Compter les états   # Les différents états : Inconnu, fluide, dense,
saturé
```

Moyenne des états en fonction de l'heure

```
Pour chaque ligne du CSV
    Grouper les lignes par heure : état

Pour chaque groupe d'état
    Sommer les états puis diviser par le nombre d'états autre qu'inconnu
```

Résultats et analyse des résultats

Zone les plus lentes

La première observation qu'il est possible de faire, c'est que certains capteurs ne sont pas actifs et ne renvoient aucun traffic, même sur une durée de 1 semaine.

```
Rte du Rhin (av.233)->Allemagne         202849
Mont??e A350 Herrenschmidt              201645
Rte du Rhin (UGC)->Ville                198736
Rte du Rhin (UGC)->Allemagne            194164
Sortie A35 -> Baggersee                 183141
Mont??e A35 Baggersee                   176882
Rte du Rhin (Danube)->Churchill         164656
Sortie A351 -> Herrenschmidt            163210
Descente Contournement Sud Rte Hopital  145813
Pont de l'Europe -> Rte du Rhin         145419
```

Ces résultats (ici uniquement le top 10) permettent de se faire une idée très rapide des zones les plus empruntées autour de Strasbourg. On peut voir d'ailleurs que le traffic est souvent sensiblement similaires entre les 2 voies, par exemple pour la route du Rhin (UGC) ou sur l'A35 au niveau de Baggersee.

Une autre remarque qui peut être faite, est que les données sont cohérentes par rapport aux expériences personnelles.

Zone les plus lentes en moyenne sur une journée

```
Rte du Rhin (av.233)->Allemagne         25356,13
Mont??e A350 Herrenschmidt              25205,63
Rte du Rhin (UGC)->Ville                24842,00
Rte du Rhin (UGC)->Allemagne            24270,50
Sortie A35 -> Baggersee                 22892,63
Mont??e A35 Baggersee                   22110,25
Rte du Rhin (Danube)->Churchill         20582,00
Sortie A351 -> Herrenschmidt            20401,25
Descente Contournement Sud Rte Hopital  18226,63
Pont de l'Europe -> Rte du Rhin         18177,38
```

On peut observer que le traffic moyen sur une journée est cohérent par rapport aux résultats précédents. Bien sûr, il s'agit évidemment du résultat attendu, mais une non corrélation de ces données permettrait de se rendre très facilement compte d'un élément exceptionnel qui s'est produit (gros rassemblement, etc...).

Somme des états par capteur

Les données des états (fluide, dense, saturé) permet de pouvoir étudier la fluidité du traffic et permet également de vérifier si les zones de gros passages, sont les zones les plus saturés du réseau (ce qui pourrait sembler intuitif).

```
Rte du Rhin (av.233)->Allemagne Inconnu : 0 / Fluide : 4336 / Dense : 75 /
Saturé : 4
Mont??e A350 Herrenschmidt  Inconnu : 0 / Fluide : 4415 / Dense : 0 / Saturé :
0
Rte du Rhin (UGC)->Ville    Inconnu : 0 / Fluide : 4400 / Dense : 14 / Saturé
: 1
Rte du Rhin (UGC)->Allemagne    Inconnu : 0 / Fluide : 4384 / Dense : 30 /
Saturé : 1
Sortie A35 -> Baggersee Inconnu : 0 / Fluide : 4008 / Dense : 387 / Saturé :
20
Mont??e A35 Baggersee   Inconnu : 0 / Fluide : 4404 / Dense : 9 / Saturé : 2
Rte du Rhin (Danube)->Churchill Inconnu : 0 / Fluide : 3775 / Dense : 523 /
Saturé : 117
Sortie A351 -> Herrenschmidt    Inconnu : 0 / Fluide : 4377 / Dense : 38 /
Saturé : 0
Descente Contournement Sud Rte Hopital  Inconnu : 0 / Fluide : 4407 / Dense :
8 / Saturé : 0
Pont de l'Europe -> Rte du Rhin Inconnu : 0 / Fluide : 4017 / Dense : 361 /
Saturé : 37
```

On peut voir que les zones les plus fréquentées ne sont pas les saturées et sont souvent bien traitées et le traffic optimisé dans ces endroits. Pour comparaison voici quelques données de comparaison :

```
Quai St.Nicolas -> Quai des P??cheurs   Inconnu : 0 / Fluide : 2800 / Dense :
1083 / Saturé : 532
RD1083 1084->1087   Inconnu : 277 / Fluide : 2616 / Dense : 1031 / Saturé :
491
Rue Alfred Kastler -> Baggersee Inconnu : 0 / Fluide : 4178 / Dense : 74 /
Saturé : 163
Schirmeck3 - Entree Ville   Inconnu : 0 / Fluide : 3970 / Dense : 161 / Saturé
: 284
All??e J.Auriol-> Mermoz    Inconnu : 0 / Fluide : 3128 / Dense : 475 / Saturé
: 812
```

Ces données permettent aussi de cibler quels sont les zones de la ville à optimiser.

Moyenne des états en fonction de l'heure

Ces données ne sont pas des plus innovantes, mais ont le mérite (tout comme pour le premier point) de confirmer l'intuition. Voici le top 10 des heures de pointes sur tout le traffic de la CUS :

```
17-28   1,1772959
17-19   1,1741072
17-46   1,1730523
17-25   1,1727215
17-22   1,1708094
17-43   1,1694373
08-42   1,164919
17-31   1,1645408
17-40   1,1639031
17-58   1,1635783
```

Ces résultats sont très cohérents par rapport au ressenti car souvent le débit semble ralenti entre 17h et 18h et vers 8h30 - 9h (mais qui est moins flagrant que le soir).

Conclusion

Bien entendu, ces analyses ne montrent qu'une partie de ce jeu de données, mais permet de s'assurer de la fiabilité des données, notamment concernant les axes les plus fréquentés.

Les possibilités sont encore grandes pour permettre de continuer d'augmenter la qualité du réseau routier de la CUS.