Technologies NoSQL

Julien Balas julien.balas@gmail.com

jan 2017 (v21)

Projet de fin de module

Dans quelques mois -votre diplôme en poche- vous avez décidé d'aller vous installer à New-York afin de trouver un emploi dans votre branche de prédilection. Confiant dans la formation que vous avez reçue, en particulier celle du module noSQL, vous voulez utiliser vos compétences pour trouver "le meilleur endroit" de la ville ou habiter.

Heureusement, la ville de New-york a un portail open data qui regorge de données qui vont vous permettre de faire ce choix https://nycopendata.socrata.com/data

Le but du projet est de télécharger quelques jeux de données, de les mettre dans une des bases noSQL vues en cours et de déterminer le meilleur quartier/arrondissement/rue ou s'installer.

Vous avez votre propre définition d'un "bon endroit", un endroit sûr, un endroit avec des commerces, des services, des restaurants, etc.

Vous ne savez pas encore ou vous allez travailler, vous n'aurez pas de voiture.

Vous aurez peut-être un vélo ou vous prendrez le métro.

Le livrable de ce projet est :

- votre définition de ce qu'est le meilleur endroit pour vivre à New-york.
- un script shell qui télécharge les jeux de données que vous allez utiliser. (au moins 2 ou 3)
- une explication du choix de la base de données que vous avez utilisé.
- un descriptif de la méthode/algorithme que vous avez implémenté pour répondre à la question
- la réponse à la question ;)
- tout le code que vous avez produit pour
  - charger les données dans la base (java, python, commande line etc.)
  - manipuler, nettoyer, ordonner les données

  - requêter la base et trouver le “meilleur endroit”
  - afficher le résultat (pas besoin d’une visualisation compliqué, une ligne de texte dans une console suffit. Si vous voulez être fancy vous pouvez)
- Votre code sera testé sur la même version d’OS que celle de vos portables (ubuntu 16.04 si ma mémoire est bonne)
- Les version de bases utilisées sont celles disponible sur les sites web de chacun des éditeurs.
- Chaque projet sera testé sur une machine vierge, n’oubliez pas de préciser comment installer les outils, modules, plugin nécessaires (pip install, pom.xml, npm install etc)

Ne mettez pas les jeux de données dans le repo. Par contre le PDF doit être dans le repo.

La date limite de réponse est le 22/02/2017 minuit, heure de mon gmail.

Envoyez moi juste le lien vers votre repository github/gitlab/bitbucket/etc. de votre projet par mail à julien.balas@gmail.com J’accuserais réception de votre mail et je clonerais le repository le 22/02/2017 à minuit.

–

Pensez à la structure des données, faut-il les transformer avant de le charger en base ? ou pas? Avez vous besoin de collections temporaires ?

Est ce que tous les calculs vont être fait dans la base? est ce qu’un petit programme externe en java/python/etc qui fait des requêtes en base peut vous aider?

Est ce que toutes les données du jeux de donnée sont intéressante ? Les vieilles données apportent elles un éclairage historique? des tendances?

Installer la base sur votre PC plutot que dans les VM des TP peut vous aider dans les manipulations.

Ne vous compliquez pas la vie outre mesure, le but est que vous trouviez quelques informations intéressante dans cet océan de données, que vous les chargiez dans une base noSQL et que vous écriviez quelques requêtes.

Par contre tout ce qui est fait pour simplifier la vie du correcteur sera apprécié à sa juste valeur.

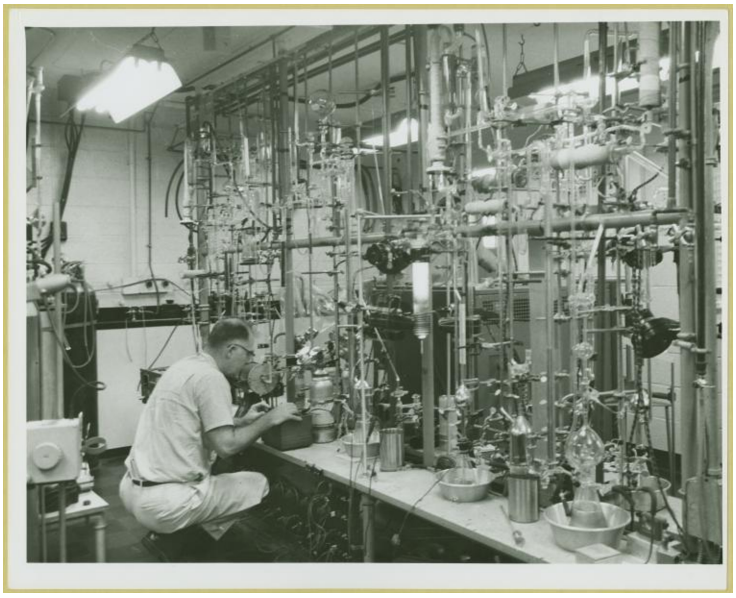

Figure 1: Un data scientist distillant de la bonne donnée