NoSQL intro

Julien Balas

nov-dec 2016 (v1)

Final project

My sister is moving to New York next month and she ask me to find here "the best" part of town to move in.

Thankfully i have a group of enthusiastic students and a large set of data available for free here : https://nycopendata.socrata.com/data

Your goal is to download some dataset, load them into a noSQL database (mongodb or neo4j) and give me the "best" borough or/and street in NY.

As everybody on earth, she is searching for good services, a safe environment, etc.

She plan to commute by bike and metro, but she don't know where she is going to work, yet.

You should provide me

- link(s) to the dataset(s) you have used
- your definition of "the best" part of town to live in
- explain the hypotheses you have made to calculate the output result.
- explain your choice of database : any of the database we have used (redis, mongodb, neo4j or cassandra)
- all the code you have written to
  - load the data into the database (java, python, commande line etc.)
  - process the data
  - query the data
  - display the data (don't be fancy on that, a list of query is ok)
- I need to be able to reproduce your result on my computer

Your work should be in a PDF for the text and in a github/gitlab/etc. repository for the code

The deadline is in 2 weeks after the last lesson –> decembre the 16th.

You are going to be free for the Christmas Holiday, yay!

Send me the link to your repository by mail, i'm going to clone them as soon as i receive the mail.

–

Think about the structure of the data, do you need to transform them before loading them into the database? or not?

Are you going to do all the processing into the database? or are you going to make a small java/python/etc program to make multiple query into the database ?

Are all the data from a dataset relevant? They may be obsolete but you may find a trend in the history of the data.

Don't over complicate things, the goal is for you to sieve some data in a large pool, get confortable with a nosql database, and write some queries.

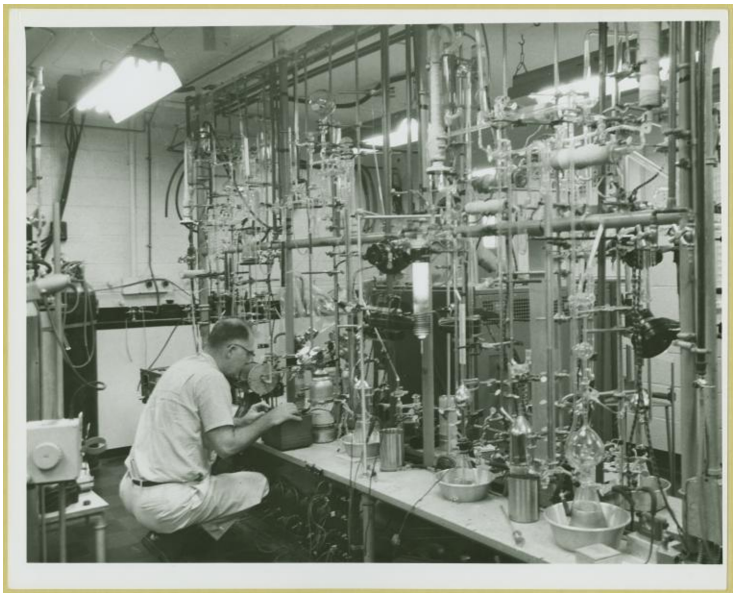

Figure 1: a data scientist hard working on data distillation