Chapitre 3 Traitement de données en tables

I Origine des données

1) Producteurs de données

Les plus gros générateurs de données sont :

- données des réseaux sociaux : likes, tweets, comments, uploads de vidéos
- données du web en général : recherche web (google trends), localisations GPS, documents, photos, vidéos
- données enregistrées par des machines : capteurs (internet des objets IOT), caméras, satellites, GPS, smartphones, appareils médicaux, appareils scientifiques.
- données générées par transactions : paiements, factures, livraisons, inventaires.

Les données informatiques sont de plus en plus nombreuses. Les fameuses "data" en anglais. Les données numériques créées dans le monde seraient passées de 1,2 zettaoctet (10^{21} octets) par an en 2010 à 1,8 zettaoctet en 2011, puis 2,8 zettaoctets en 2012 et s'élèveront à 40 zettaoctets en 2020. À titre d'exemple, Twitter générait en janvier 2013, 7 téraoctets (10^{12}) de données chaque jour et Facebook 10 téraoctets. En 2014, Facebook Hive générait 4 000 To de data par jour.
Ce sont les installations technico-scientifiques (météorologie, satellites, capteurs) qui produiraient le plus de données. De nombreux projets de dimension pharaonique sont en cours. Le radiotélescope “Square Kilometre Array” par exemple produira 50 téraoctets de données analysées par jour, tirées de données brutes produites à un rythme de 7 000 téraoctets par seconde.

Chaque année, l’humanité produit environ un volume d’informations numériques de l’ordre de la centaine de zettaoctet (1 zettaoctet = 10^{21} octets = l'équivalent de 250 milliards de DVD)

1 Méga = 10^6 octets , 1 Giga = 10^9 octets, 1 Téra = 10^{12} octets, 1 Péta = 10^{15} octets, 1 Exa = 10^{18} octets, 1 Zéta = 10^{21} octets

2) Format des données produites

On ne peut tous les citer. Les principaux sont :

- **le format JSON (Javascript Object Notation)** : voici un exemple d’un enregistrement d’un tweet au format JSON (voir page suivante). Ce format est très similaire à celui d’un dictionnaire en python avec des clés et des valeurs.
- **le format xml** : il ressemble à de l’HTML avec des balises (voir page suivante)
- **le format CSV** (comma separated values) : une ligne avec des descripteurs séparées par des virgules puis des données séparées par des virgules ou des points virgules (voir page suivante).

Ces formats (un peu vieux) sont encore employés à l’heure actuelle mais d’autres plus adaptés se développent comme AVRO et PARQUET ou une représentation sous forme de graphes avec RDF.

Utilisations

1) Dans le tweet indiqué, tweet.location renvoie "Nantes"
2) Dans le tweet indiqué, tweet.status.favorite_count renvoie 2
3) Un objet JSON se construit ainsi : losc = {" sport " : football, " ville " : Lille, " classement " : 4}

Format JSON :

Nous donnons dans cette section à la figure A2.2 un extrait de biographie du compte Twitter du Château de Nantes collecté après la *#MuseumWeek* (le samedi 4 avril).


```
{"created_at": "Thu May 05 14:28:44 +0000 2011",
 "id": 293534469,
 "id_str": "293534469",
 "name": "Château de Nantes",
 "screen_name": "ChateauNantes",
 "location": "Nantes",
 "url": "http://t.co/AlNzCLd5AX",
 "description": "Bienvenue au Château des ducs de Bretagne - Musée d'histoire urbaine de la ville de Nantes / LT #expoLaboureur",
 "protected": false,
 "followers_count": 3907,
 "friends_count": 317,
 "listed_count": 157,
 "favourites_count": 452,
 "geo_enabled": true,
 "verified": false,
 "statuses_count": 2813,
 "status": {
      "contributors": null,
      "truncated": false,
      "text": "RT @LucasLechat: Au pied du plus haut clocheton du @ChateauNantes http://t.co/LH4bU50wvd",
      "created_at": "Sat Apr 04 15:05:23 +0000 2015"
      "geo": null,
      "coordinates": null,
      "place": null,
      "contributors": null,
      "retweet_count": 1,
      "favorite_count": 2,
      "entities": {
           "hashtags": [
                {"text": "AnnedeBretagne", "indices": [50,65] },
                 ...
                {"text": "architectureMW", "indices": [101, 109] }],
           ... },
      "favorited": false,
      "retweeted": false,
      "lang": "fr" }
...
}
```


Figure A2.2 | Extrait de biographie (id 293534469) collectée durant la *#MuseumWeek* (2015) ; biographie du Château de Nantes. Les attributs de métadonnées primaires sont indiqués en gras, les autres non. Les valeurs associées aux attributs (primaires ou non) sont indiquées en gris.

Format xml :


```xml
<?xml version="1.0" encoding="UTF-8"?>
<liste>
    <personne>
        <NOM>Perrin</NOM>
        <PRENOM>Léo</PRENOM>
        <AGE>18</AGE>
    </personne>
    <personne>
        <NOM>Petit</NOM>
        <PRENOM>Loic</PRENOM>
        <AGE>32</AGE>
    </personne>
    <personne>
        <NOM>Leroux</NOM>
        <PRENOM>Pierre</PRENOM>
        <AGE>27</AGE>
    </personne>
</liste>
```


Format CSV :


```
Nom,Prenom,Age
Perrin,Léo,18
Petit,Loic,32
Leroux,Pierre,27
```