NF26 - Stockage en haute volumétrie et applications - Projet P21

1 Introduction

Dans le cadre du projet en NF26, j'ai étudié le jeu de données Safecast Radiation Measurements comportant 82282963 mesures réalisées entre 1969 et 2020. Mon étude contient 1265248 mesures, hormis pour la question 3 qui en comporte 6611632.

L'analyse du jeu de données est réalisée par le biais de Cassandra pour les questions 1 et 2, et Spark pour la question 3. Nous expliquerons le choix des modèles et des outils dans la section 2.

Le jeu de données étudié comporte des mesures avec comme unité le cpm (count per minute), µsv ou encore status mais ces derniers sont très faiblement représentés comparés au cpm (de l'ordre de 4% dans le jeu de données total et 3.10^{-6}% dans les mesures étudiées). Nous nous intéressons donc seulement aux mesures dont l'unité est le cpm. D'après [1], 350 cpm est converti en 1 µSv/h qui correspond au Sievert, l'unité utilisée pour donner une évaluation de l'impact des rayonnements sur l'homme.

2 Modèle

Pour les questions 1 et 2, nous avons utilisé Cassandra pour créer deux tables. L'une contenant toutes les données pour les préserver au cas où le fichier csv ne serait plus disponible, et l'autre pour contenir les zones les plus radioactives. L'accès aux mesures peut donc se faire sur un nombre limité de partitions en récupérant la distribution des mesures ainsi que les zones les plus radioactives. En créant seulement une table qui regroupe la localisation des mesures effectuées, on limite le nombre de colonnes à affecter à la table et les partitions sont donc réduites.

Pour la question 3, nous avons pris plus de mesures pour avoir une plus grande quantité à travers les années. Ainsi, en utilisant Spark, nous pouvons directement accéder aux mesures selon les années sans créer de tables sur Cassandra. L'analyse des résultats peut donc se faire en quelques lignes de code et est plus rapide car ne nécessite pas l'insertion des données.

3 Analyse des requêtes

1. La Figure 1 représente la distribution des mesures en cpm relevées. On observe un pic de fréquence pour les valeurs situées entre 35 et 40 cpm, ce qui correspond à environ 0.1 µSv/h. A titre de comparaison, la dose de radiations absorbée lorsqu'une personne mange une banane est de 1 µSv [2]. Néanmoins, il y a des valeurs dépassant les 14 µSv/h et atteignant même jusqu'à 65 µSv/h (=569.4 mSv en une année), sachant que la limite annuelle des ouvriers du nucléaire est fixée à 50mSv. Nous pouvons donc en déduire qu'il y a eu au cours des années des évènements ayant contribué à fortement augmenter la radiation comme l'accident nucléaire de Fukushima en 2011.

2. La Figure 2 [1] représente la localisation des mesures relevées reportées sur un planisphère. On y observe quatre zones géographiques les plus radioactives sur les mesures étudiées. Pour distinguer laquelle des zones est la plus radioactive entre l'Europe et l'Océanie, nous avons calculé la somme des valeurs dans ces zones et nous trouvons respectivement 303740 et 603851 cpm. *Nous concluons donc que les trois zones géographiques les plus radioactives sur les mesures étudiées sont : l'Amérique du Nord, le Japon et l'Océanie.* Ces résultats étaient prévisibles car les Etats-Unis sont la première puissance nucléaire et effectuent de nombreux tests nucléaires dans leur région. D'après [3], les pays comme les Etats-Unis, la France ou le Royaume-Uni ont conduit beaucoup de tests nucléaires entre les années 1960 et 1996 dans les îles de l'Océan Pacifique Sud, près de l'Australie, notamment en Polynésie française par exemple. Quant au Japon, les accidents nucléaires comme à Fukushima en 2011 expliquent en grande partie la très forte radioactivité dans cette zone. Hormis la visualisation sur une carte, nous avons aussi relevé trois points correspondant au centre des zones les plus radioactives. En effet, au vu de l'apparente séparation des points sur la carte, nous avons jugé raisonnable de calculer les centroïdes pour trouver les zones les plus radioactives. Ces points ne correspondent pas explicitement aux zones les plus radioactives mais constituent le centre de ces zones. Pour ce faire, nous avons utilisé l'algorithme des KMeans en streaming sur nos longitudes et latitudes. Il en résulte trois points dont les coordonnées sont présentés dans la Table 1[2]. On peut voir qu'ils correspondent effectivement au Japon, l'Océanie et l'Amérique du Nord.

3. Dans le jeu de données, il existe des radiations mesurées lors d'années supérieures à 2020. Nous ne prenons en compte seulement les années jusqu'à 2020. En fait, plus de 99.99% des radiations ont été mesurées entre les années 2011 et 2017. La Table 2 référence le nombre de radiations mesurées pour chaque année (après avoir enlevé les données aberrantes où les dates n'étaient pas identifiables). Nous pouvons donc seulement évaluer les mesures entre ces années. Par conséquent, il nous est impossible de réaliser une analyse de cycles sur plusieurs décennies. La Figure 3 a été

[1] Cette figure est de moindre qualité car le rendu vectoriel était trop lourd, j'ai donc dû utiliser un format png.

[2] Ces coordonnées ne sont pas mis à l'échelle contrairement aux points représentés sur la carte de la Figure 2.

réalisée en prenant un sous-échantillon aléatoire du jeu de données. L'année 2016 semble présenter les radiations les plus élevées. Or, d'après le relevé de l'activité solaire [4], en l'année 2016, il y a eu 16 éruptions solaires de catégorie M et les autres sont classées C ou moins [5]. En comparaison, l'année 2015 comptabilise deux éruptions solaires classées X et plus de 50 de classées M. Or, on observe une valeur moyenne plus basse pour l'année 2015 que pour l'année 2016. De même, l'année 2013 comptabilise 12 éruptions solaires classées X et plus d'une quarantaine classées M. Pourtant, la moyenne des mesures en 2013 est bien inférieure à celle de 2016. *On peut donc en conclure qu'il n'y a pas de lien avec les éruptions solaires.* Enfin, on peut également remarquer qu'il semble exister des valeurs aberrantes sur les valeurs de radiations mesurées, ce qui expliquerait les fortes incertitudes pour l'année 2014.

4 Conclusion

L'étude de ce jeu de données nous a permis d'explorer les outils d'analyse tels que Cassandra ou Spark. Nous avons pu tirer plusieurs observations sur les mesures de radiations comme leur distribution, les zones les plus radioactives ou encore l'aspect temporel. Pour ce dernier point, nous avons été limité en données car celles-ci ne s'étendent pas sur un spectre d'années assez large. Il faut donc considérer la conclusion sur l'éruption solaire comme étant discutable dans la mesure où plus de données au fil du temps nous aurait permis d'établir une étude plus précise en réalisant un test de constance ou une analyse de cycles par exemple. Enfin, ce projet nous a également donné l'opportunité d'essayer le calcul en streaming via l'utilisation des KMeans.

Annexes

A Requête 1

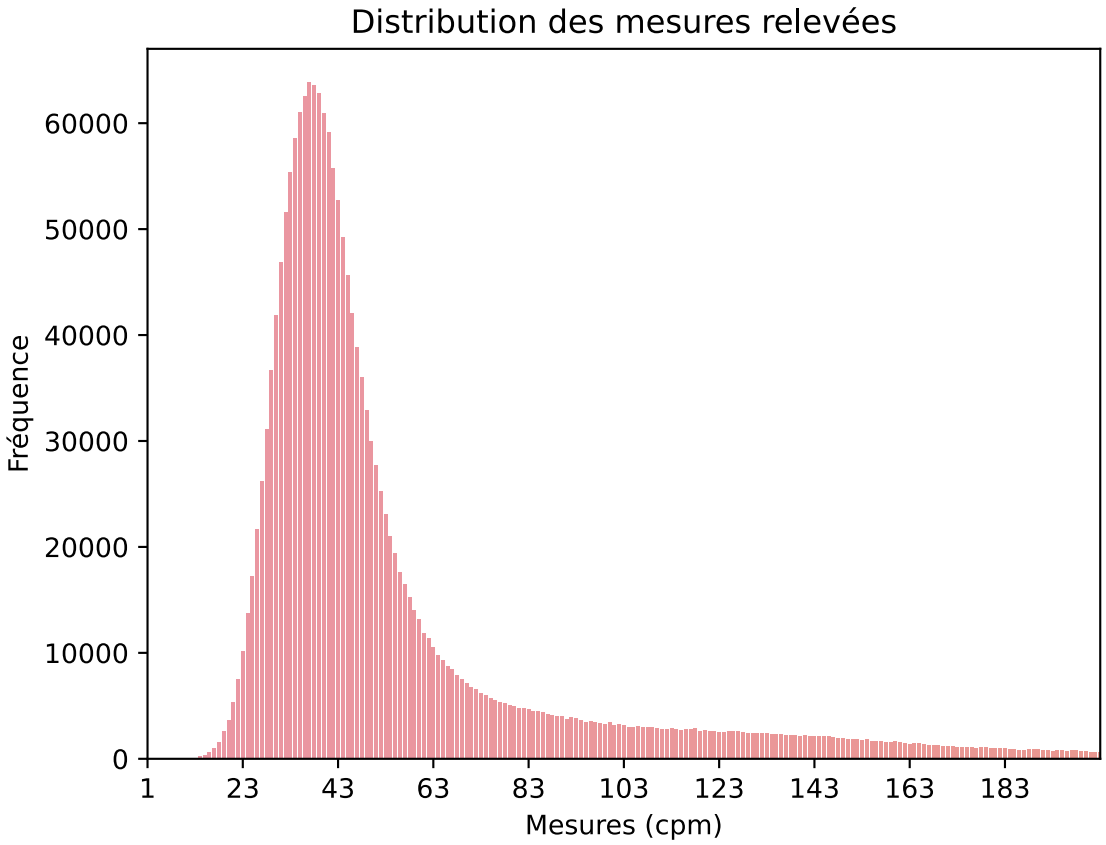

Figure 1: Distribution des mesures relevées

B Requête 2

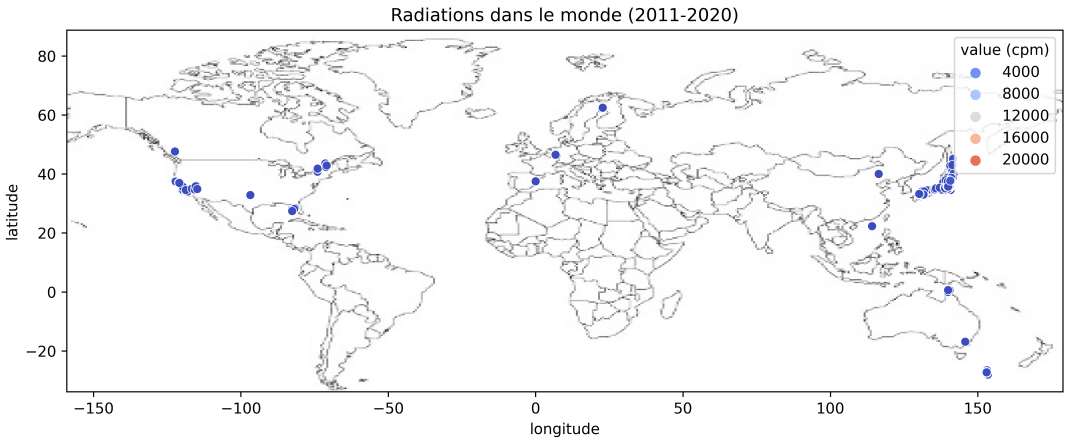

Figure 2: Radiations dans le monde (2011-2020)

Table 1: Centroïde des zones les plus radioactives

Centroïde	longitude	latitude
1	32.9	130.0
2	39.1	-80.5
3	-27.2	153.0

C Requête 3

Table 2: Nombre de radiations mesurées par année

Année	Nombre de mesures
2004	1
2010	1
1983	3
2000	3
2019	8
2003	24
2020	28
1969	89
1981	130
2006	342
1970	2028
2011	2118153
2012	3600216
2013	10124683
2014	11631480
2015	15918111
2016	17833096
2017	21405526

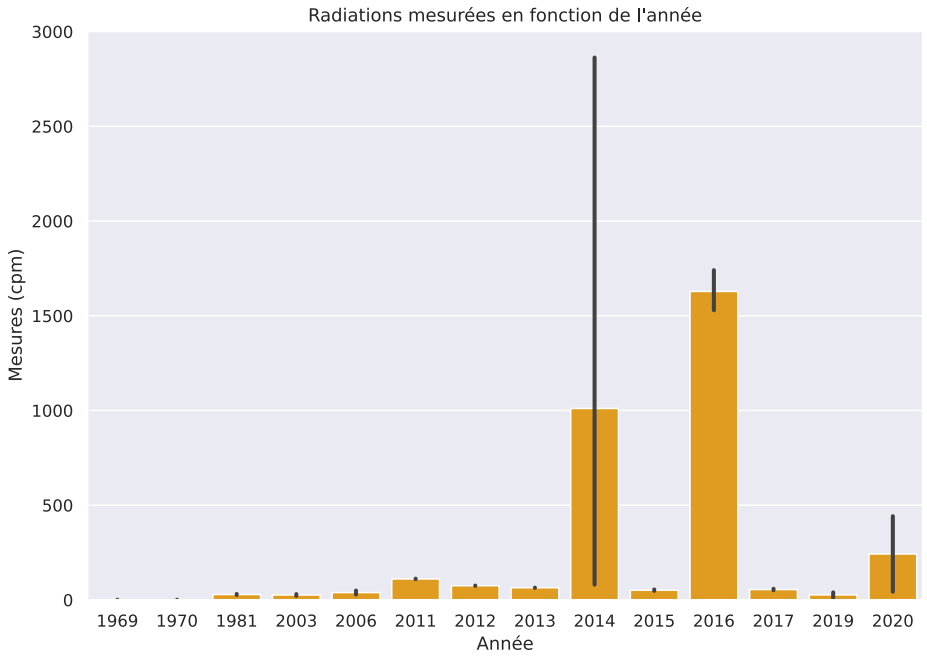

Figure 3: Radiations mesurées en fonction de l'année

References

[1] Safecast. Methodology. *Disponible en ligne à cette URL.*

[2] Randall Munroe. Radiation Dose Chart. *Disponible en ligne à cette URL.*

[3] Aleksandar Novaković. Death in paradise: The aftermath of nuclear testing in Australia and Oceania. *Disponible en ligne à cette URL.*

[4] Space Weather. Top 50 solar flares of the year 2020. *Disponible en ligne à cette URL.*

[5] Space Weather. Solar Flare Classifications. *Disponible en ligne à cette URL.*