Analyse exhaustive sur les vols aux états-unis sur une jeu de données de 2015

Ghouibi Ghassen

Université Paris 8

ghassen.ghouibi@etud.univ-paris8.fr

20/01/2020

Table des matières

1. Introduction
   - Définition
   - Problématique
   - Sujet
2. Traitement de données
   - Hive importer les fichiers
   - Hive exporter les fichiers
   - Python — hive
   - Résultat
3. Exemple de résultat obtenu
4. Conclusion
   - Sources
   - Questions

Introduction

Introduction

C'est quoi la différence entre une architecture centralisée et une architecture distribuée ?

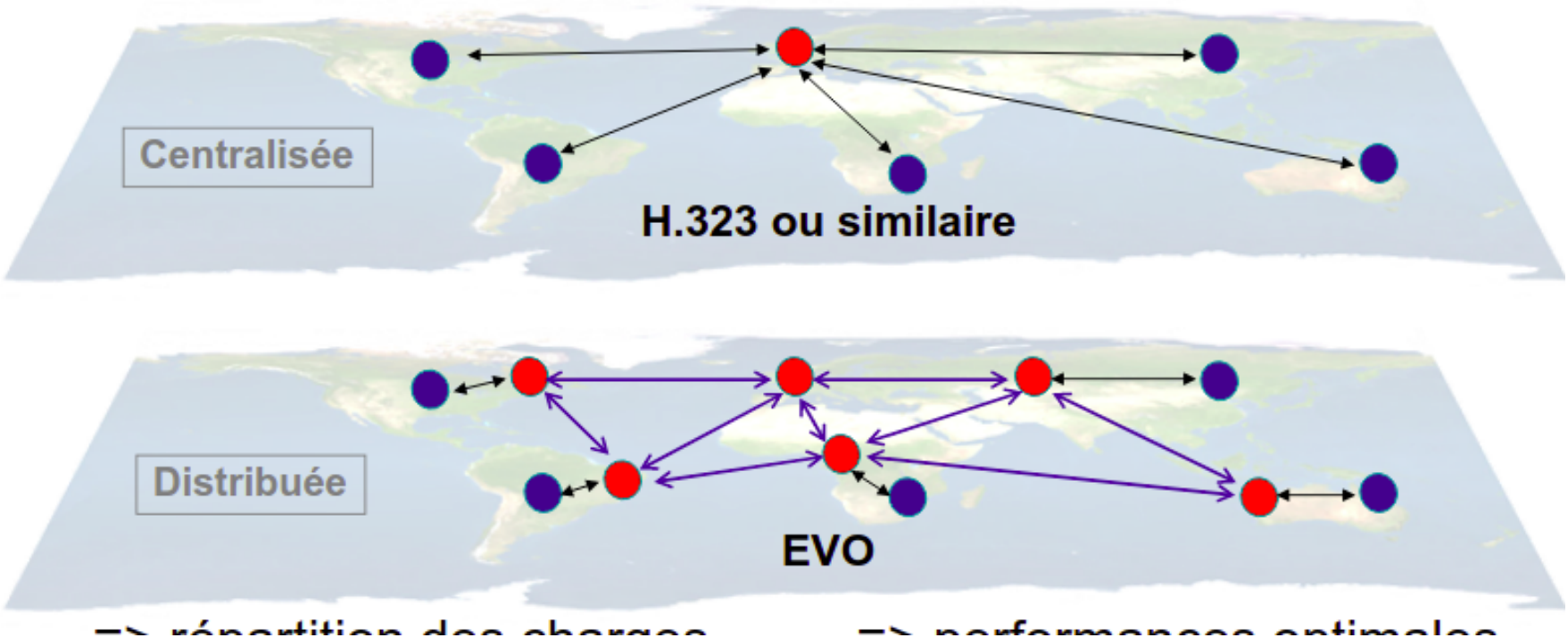


Figure: Les architectures distribuée et les architectures centralisée

Introduction — Problématique

- Est ce que l'aéroport peut avoir une influence sur le retard des vols ?

Introduction — Sujet

Dans cette présentation on étudier quelques étapes pour la réalisation du projet qui sera comment peut on se fier au aéroport pour prédire les retards.

EMR — Hive importer les fichiers

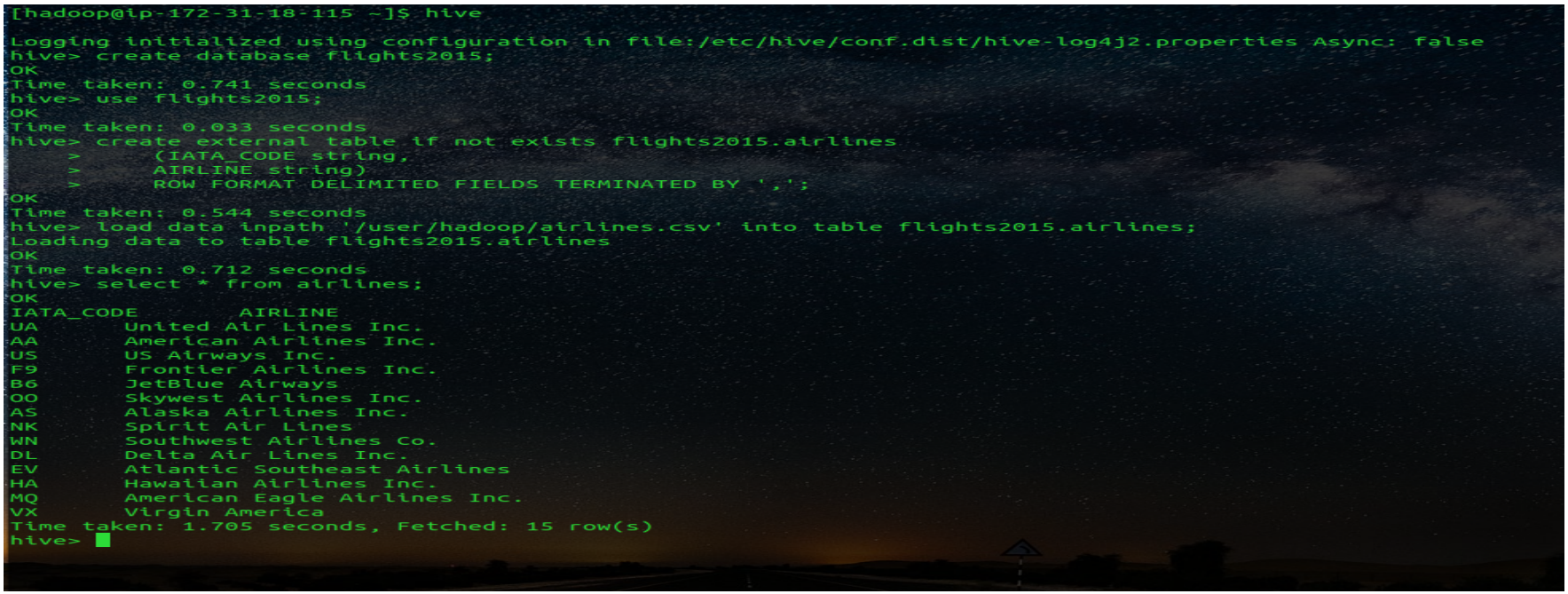


Figure: Importer les fichiers et les stocker dans des tables

EMR — Hive exporter les fichiers

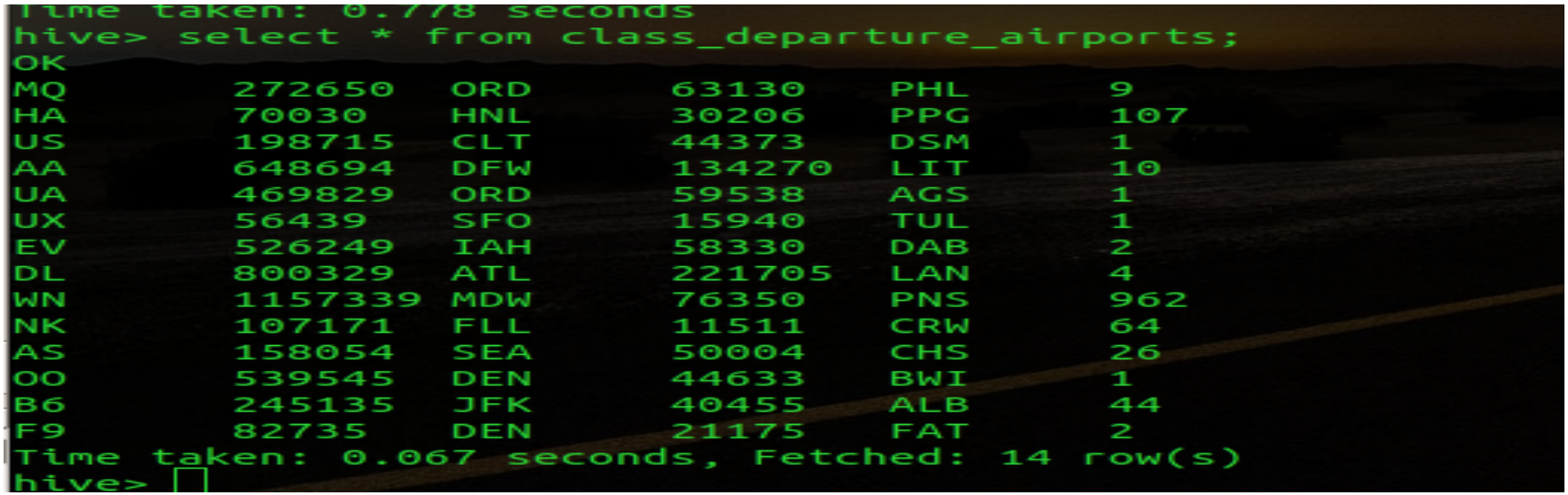


Figure: Table du résultat

```
insert overwrite local directory '/home/hadoop/' row
format delimited fields terminated by ',' select * from
class_departure_airports; hadoop dfs -copyFromLocal
/home/hadoop/ s3://flights2015/input/
```

Python — réprésentation de table

```
flights={}
airports_best={}
airports_worst={}

with open('file.csv', newline='') as csvfile:
    reader = csv.DictReader(csvfile)
    for row in reader:
        print(row['IATA_CODE'], row['NUMBER_OF_FLIGHTS'])
        flights.update( {row['IATA_CODE'] :int(row['NUMBER_OF_FLIGHTS'])} )
        airports_best.update( {row['MOST_LANDED_AIRPORT'] :int(row['NB_OF_MLA'])} )
        airports_worst.update( {row['LESS_LANDED_AIRPORT'] :int(row['NB_OF_LLA'])} )



plt.figure(1)
plt.bar(range(len(flights)), list(flights.values()), align='center')
plt.xticks(range(len(flights)), list(flights.keys()))
plt.xlabel("Nom de compagnies aériennes")
plt.ylabel("Nombre de vol par an")
plt.title("Classement des compagnies aériennes par nombre de vol")

plt.figure(2)
plt.bar(range(len(airports_best)), list(airports_best.values()), align='center')
plt.xticks(range(len(airports_best)), list(airports_best.keys()))
plt.xlabel("Nom de l'aéroport")
plt.ylabel("Nombre de vol par an")
plt.title("Classement des aéroports avec le plus d'influence par nombre de vol")

plt.figure(3)
plt.bar(range(len(airports_worst)), list(airports_worst.values()), align='center')
plt.xticks(range(len(airports_worst)), list(airports_worst.keys()))
plt.xlabel("Nom de l'aéroport")
plt.ylabel("Nombre de vol par an")
plt.title("Classement des aéroports avec le moins d'influence par nombre de vol")

plt.show()
```

Figure: Code python pour la réprésentation graphique

Data Visualisation

La visualisation des données est un ensemble de méthodes de représentation graphique, en deux ou trois dimensions, utilisant ou non de la couleur et des trames.
Les moyens informatiques permettent de représenter des ensembles complexes de données, de manière plus simple, didactique et pédagogique.

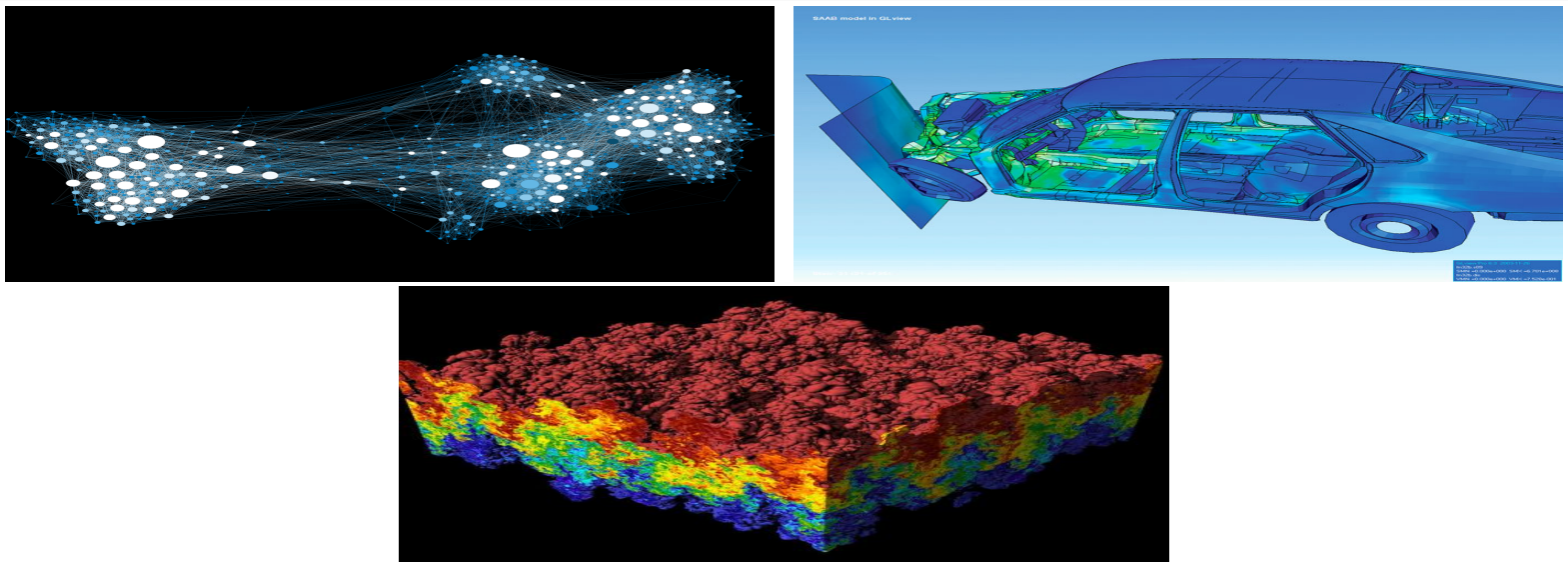
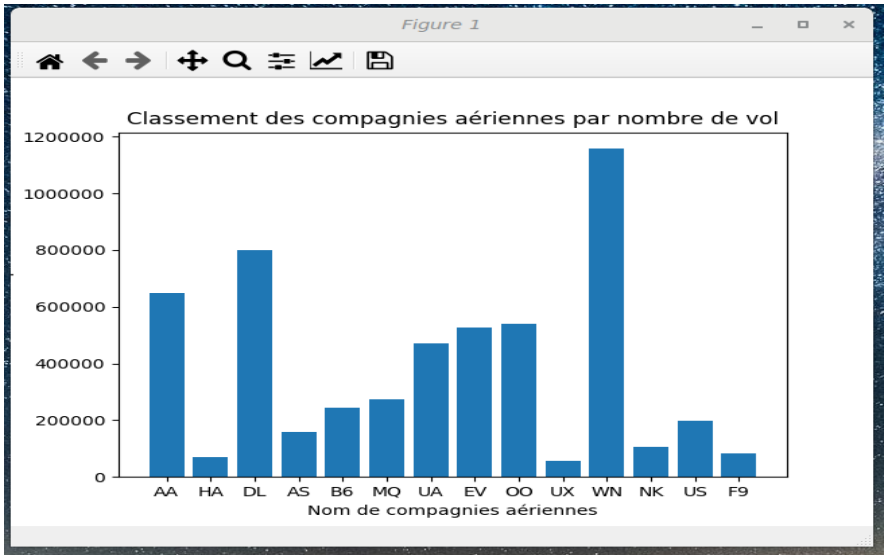
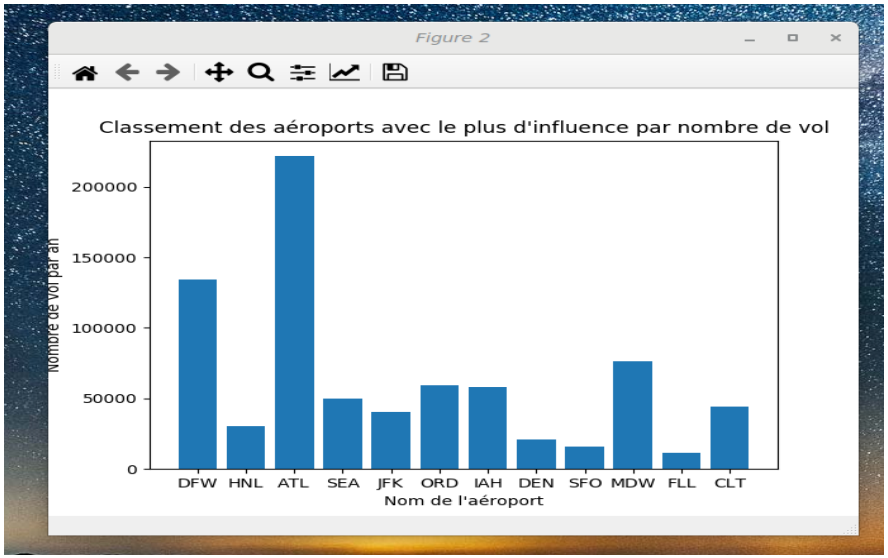


Figure: Exemple différents types de visualisation

Exemple de résultat obtenu





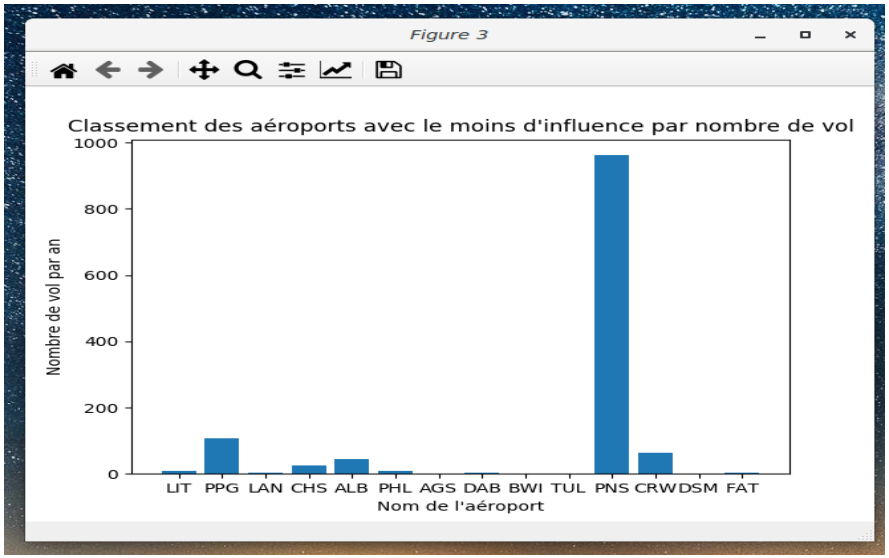


Figure: Résultat obtenu

Conclusion

Conclusion

Les résultats obtenus suite à ce projet montre les classements des aéroports et des compagnies aériennes sur des masses de données d'où l'importance des architecture distribuées pour effectuer ce genre d'opération qui sera un outils incourtournable pour les Data Engineer et qu'on ne peut pas produire avec une seule machine.

Conclusion - Sources


https://www.rosettahub.com/welcome


https://www.kaggle.com/usdot/flight-delays


https://aws.amazon.com/fr/


https://spark.apache.org/


https://www.youtube.com/watch?v=lZvs-YNk4V0


https://stackoverflow.com/

Conclusion - Fin

Merci pour votre attention !