# **MD4 - Big Data Architecture - Projet Final**

|  |  |
| --- | --- |
| Version | Changements |
| 1.0 | Document initial |
| 1.1 | Inviter @AJRdev avec le rôle maintainer |

## Objectifs



### Exploiter la data de la SNCF

La structure de dossier pour cette objectif :

* un dossier **sncf-features** contenant les notebooks et datasets
* un dossier **data** dans le dossier **sncf-features** contenant les datasets
* un notebook nommé **sncf-features.ipynb** dans le dossier **sncf-features**
* un fichier **README.md** dans le dossier **sncf-features** listant les sources de vos datasets

Sources des datasets : <https://ressources.data.sncf.com/explore>

1. A l'aide des datasets SNCF de votre choix :

* créer une fonction **itineraryPrice(station1, station2)** qui estimera le prix du trajet en **1ère classe** entre les 2 noms de gares passés en paramètres
* cette fonction devra utiliser PySpark pour exploiter les datasets choisis
* lister **4 résultats** d'appels de cette fonction avec des gares différentes pour chaque appel
* la fonction devra retourner **None** dans le cas où :
  + les gares passés en paramètre n'existent pas
  + le prix pour le trajet est inexistant

2. Etablissez à l'aide du module python [geopy](https://pypi.org/project/geopy/) un classement descendant des distances pour chaque trajets possible

3. Créer un **Dataframe** listant tous les trajets possibles avec 3 colonnes : Trajet, Distance, Prix en 1ère classe

4. Créer une représentation graphique mettant en avant la corrélation entre le prix et la distance pour chaque trajet possible

### Effectuer une analyse exploratoire libre

La structure de dossier pour cette objectif :

* un dossier **free-analysis** contenant les notebooks et datasets
* un dossier **data** dans le dossier **free-analysis** contenant les datasets
* un notebook nommé **free-analysis.ipynb** dans le dossier **free-analysis**
* un fichier **README.md** dans le dossier **free-analysis** listant les sources de vos datasets

1. A l'aide de 2 datasets (ou plus) de votre choix (provenant de Kaggle ou autres sources), établissez 3 analyses permettant de mettre en avant des corrélations entre les différents datasets

2. Chaque analyse devra :

* être faite à l'aide de PySpark avec soit des RDD ou des DataFrames
* comporter une représentation graphique de votre choix

**Notes :**

* Vous être libre de choisir n'importe quel librairie python pour enrichir votre analyse

### 

## Livraison

* Créer un répertoire de projet Gitlab nommé **md4-big-data-final** par groupe
* Créer un **README.md** listant les différents membres du groupe
* Inviter **@AJRdev** sur ce répertoire avec le rôle **Maintainer**

## 

## Evolutions

* Estimer le prix d’un trajet entre 2 lieux avec la fonction **itineraryPrice**