

Big Data. TD 3.


Sergey Kirgizov

ESIREM


Apache Spark

L'objectif est de se familiariser avec . Spark nous permet d'utiliser l'un des langages de programmation suivants : Scala, Python, Java, R.

 **EXERCICE 1 : Télécharger Apache Spark**
<https://spark.apache.org/downloads.html>

 **EXERCICE 2 : Installer Spark et lancer l'exemple du calcul du nombre π**
<https://spark.apache.org/docs/latest/>

```
./bin/spark-submit examples/src/main/python/pi.py
```


 **EXERCICE 3 : Suivre le tutoriel officiel de démarrage rapide**
<https://spark.apache.org/docs/latest/quick-start.html>


 **ASTUCE : Pour afficher le contenu d'un fichier dans le console python du Spark on peut faire**


```
data = spark.read.text('file.txt')
data.foreach(print)
# ou bien
data.show()
```

 **ASTUCE : Pour afficher la documentation d'une fonction f en python on peut faire**

```
help(f)
```

 **EXERCICE 4 : Préparer un fichier de taille 1M lignes. Vous pouvez utiliser soit**
— le fichier <https://kirgizov.link/teaching/esirem/bigdata-2021/dataset/wikirank-fr.tsv.gz>
— le générateur <http://www.ordinal.com/gensort.html>

 **EXERCICE 5 : Trier les fichiers de tailles différentes avec Spark tout en mesurant le temps du calcul. Comparer ce temps du calcul avec le temps pris par le logiciel classique d'unix :**
time sort file.tsv

 **EXERCICE 6 : Lire le fichier wikirank-fr.tsv à l'aide de la fonction spark.read.csv**



EXERCICE 7 : Lire la documentation de Spark SQL
<https://spark.apache.org/docs/latest/sql-getting-started.html>.

Afficher le TOP 20 des articles les plus populaires.



EXERCICE 8 : Calculer les corrélations entre le nombre d'auteurs et la popularité des articles.



EXERCICE★ 9 : Télécharger le jeu de données complet du WikiRank.
Faire un gros fichier `all_languages.tsv` en utilisant par exemple le logiciel `cat` d'unix.
Trier le fichier avec Spark tout en mesurant le temps du calcul. Comparer ce temps du calcul avec le temps pris par le logiciel classique d'unix `sort`.
https://figshare.com/articles/WikiRank_05_2019_-_quality_scores_popularity_and_AI_for_Wikipedia_articles/8231273/2