

## ITÉRATION 2

# Premières manipulation de données avec Spark

### Modalités

- Travail individuel en autonomie
- 1,5 journées en présentiel

### Livrables

- Code terminé

### Objectifs

- Prendre en main le fonctionnement de Spark
  - Découverte avec les listes
  - Exploration avec les textes

### Compétences

- Je manipule des données avec PySpark

## 1.1 – Configuration de votre environnement

1h – Présentiel

Ça y est, vous avez installé votre environnement de travail, il est temps maintenant d'explorer les différentes choses que vous pouvez faire avec PySpark !

En creusant sur différents sites vous comprenez qu'il est possible de configurer votre environnement Spark avec `SparkContext` et `SparkConf`. Vous vous plongez dans la doc pour mettre en place votre environnement.

Au fil de vos lectures vous comprenez de mieux en mieux l'intérêt de Spark et de la facilité avec laquelle il permet de paralléliser les calculs!

### RESSOURCES

- Documentation sur les public classes :  
<https://spark.apache.org/docs/latest/api/python/reference/pyspark.html#public-classes>
- Documentation sur le Spark context :  
[https://www.tutorialspoint.com/pyspark/pyspark\\_sparkcontext.htm](https://www.tutorialspoint.com/pyspark/pyspark_sparkcontext.htm)  
<https://spark.apache.org/docs/latest/api/python/reference/pyspark.html#spark-context-apis>

### COMPÉTENCES ASSOCIÉES

- Configuration d'un environnement (Py)Spark

## 1.2 – Découverte de Spark avec les listes

5h – Présentiel

Maintenant que vous avez mis en place votre environnement, vous décidez de commencer à explorer les différentes fonctionnalités de Spark sur des listes. Plusieurs questions vous viennent en tête au fil de vos recherches documentaires.

Vous commencez par créer une liste très simple pour tester les différentes fonctions :

```
l = [1, 2, 3, 4, 5]
```

Vous vous demandez quelles manipulations vous pouvez faire dessus en vous remémorant vos premiers cours de Python...

- Comment distribuer ma liste dans un RDD ? ... et d'ailleurs, c'est quoi un RDD d'abord ?!
- Maintenant que je vois à peu près ce que c'est, comment est-ce que je récupère le contenu des RDDs ? ... elle fait quoi la méthode `'glom()'` ?
- ... Spark *lazy evaluation*: qu'est-ce que ce truc ?! Que se passe-t'il quand je lance un `collect()` ?

- quels types de manipulation puis-je faire avec `map` ? `reduce` ? `saveAsTextFile` ? Les autres trouvées dans ma CheatSheet (cf ressources ci-dessous) ?

Vous vous dites qu'elle est bien belle cette petite liste, mais qu'il faut passer aux choses sérieuses maintenant ! Vous augmentez le volume de vos données et continuez d'explorer :

```
1 = [1, 2, 3, 4, 5]*1000000
```

En ayant à présent assez d'explorer sans but précis, vous vous fixez un petit challenge ! Vous prenez les données des arbres remarquables de Paris et répondez à plusieurs questions que vous vous posez :

- Quel âge a notre plus vieil ami ?
- Quel est le volume du plus grand arbre ?
- Quelle est la hauteur moyenne de ces spécimens ?
- Dans quel arrondissement y a-t-il le plus d'arbres ?
- Combien y a-t-il d'arbres au Cimetière du Père Lachaise ?
- ...

## RESSOURCES

- Documentation sur les RDD :  
<https://spark.apache.org/docs/latest/api/python/reference/pyspark.html#rdd-apis>
- Lazy evaluation :  
<https://towardsdatascience.com/3-reasons-why-sparks-lazy-evaluation-is-useful-ed06e27360c4/>
- Cheat Sheet PySpark :  
[https://images.datacamp.com/image/upload/v1676303379/Marketing/Blog/PySpark\\_RDD\\_Cheat\\_Sheet.pdf](https://images.datacamp.com/image/upload/v1676303379/Marketing/Blog/PySpark_RDD_Cheat_Sheet.pdf)
- Données sur les arbres de Paris :  
<https://opendata.paris.fr/explore/dataset/arbresremarquablesparis/information/>
- Quelques pointeurs supplémentaires:  
<https://spark.apache.org/docs/latest/sql-getting-started.html>

## COMPÉTENCES ASSOCIÉES

- Manipulation de listes avec Spark

5h – Présentiel

## 1.3 – Manipulation de texte avec Spark

Maintenant à l'aise avec les listes, vous vous lancez dans les choses sérieuses, après tout c'est sur du texte que vous avez besoin de travailler ! Vous choisissez un livre dans le projet Gutenberg et commencez à tester quelques manipulations découvertes précédemment.

Vous décidez de vous fixer de nouveaux challenges ! Vous prenez le livre “Beautiful Stories” de Shakespeare, plusieurs questions vous viennent naturellement :

- Quel est le thème qui ressort le plus dans ces belles histoires ? Vous regardez les mots les plus utilisés afin d'avoir un aperçu du champ lexical de ces histoires.
- Vous avez découvert une superbe histoire que vous voulez partager avec un ami, vous vous lancez dans la découpe des différentes histoires pour les exporter séparément.

### RESSOURCES

- Gutenberg Project:  
<https://www.gutenberg.org/>  
<https://www.gutenberg.org/ebooks/author/65>  
<https://www.gutenberg.org/cache/epub/1430/pg1430.txt>

### COMPÉTENCES ASSOCIÉES

- Manipulation de textes avec Spark

### Livrables

Manipulation du texte choisi :

- ➔ Résultats des challenges que vous vous êtes fixés