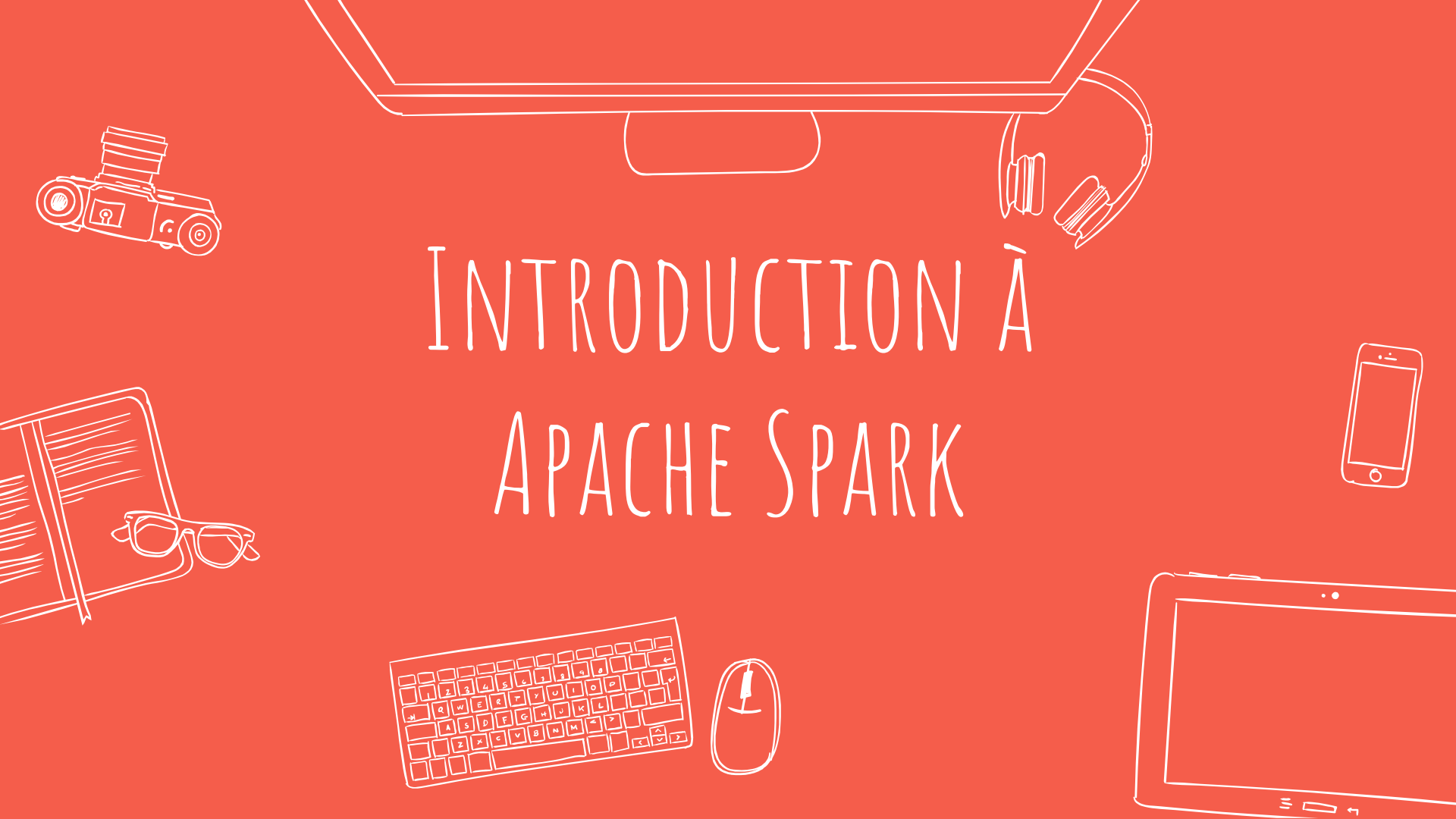
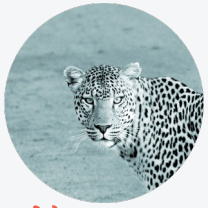

Introduction à
Apache Spark



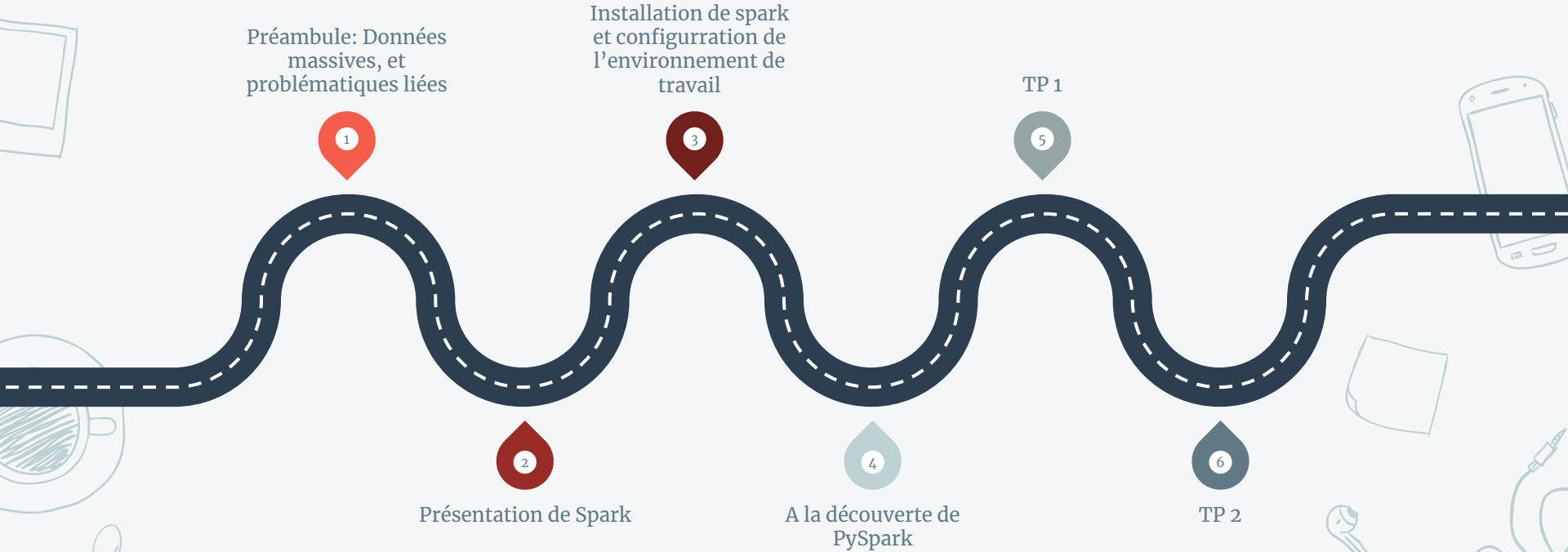
HELLO!

F. J. AMON D'ABY

Ingénieur d'Etat en Informatique

Software Development, Software Architecture, AI, Big Data

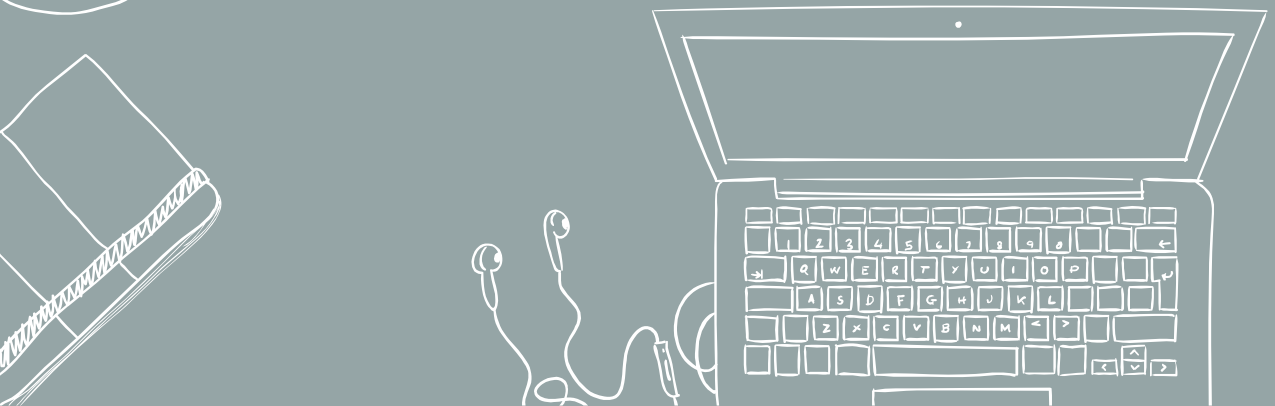
Roadmap


Préambule: Données massives, et problématiques liées
1
Présentation de Spark
2
Installation de spark et configurration de l'environnement de travail
3
A la découverte de PySpark
4
TP 1
5
TP 2
6

1.

Préambule

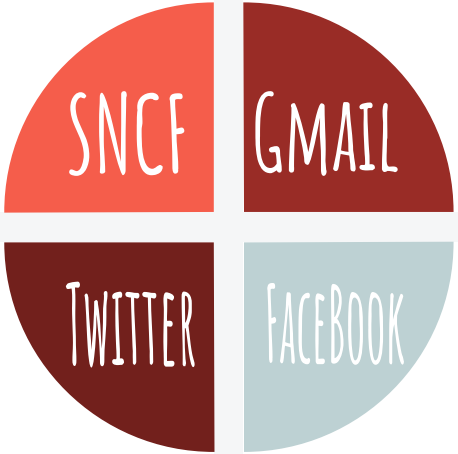
Vous avez dit données...



Volume, Variété, Vélocité

Voyages-sncf.com

11 millions de visiteurs uniques par mois



Gmail

425 Millions d'utilisateur (fin 2012)

284 Millions d'utilisateurs actifs

Twitter

890 Millions d'utilisateurs actifs par jour en moyenne (fin 2014)

Facebook

Mégadonnées (Big Data) =

- des données qui :
  - sont trop volumineuses
  - ou ayant une arrivée trop rapide
  - ou une variété trop grande
- pour :
  - permettre de les ranger directement dans des bases de données traditionnelles (Relationnelles)
  - ou de les traiter par les algorithmes actuels [1].
- et qui nécessitent l'utilisation d'outils spécifiques
  - La figure suivante présente différents outils qui ont été développés pour traiter du big data. Il en existe un grand nombre.

[1] MOTHE J., PITARCH Y., GAUSSIER E. Big Data: Le cas des systèmes d'information, Revue Ingénierie des Systèmes d'Information, Hermès Editeur, Vol. 19/3, (2014).

Introduction à Apache Spark

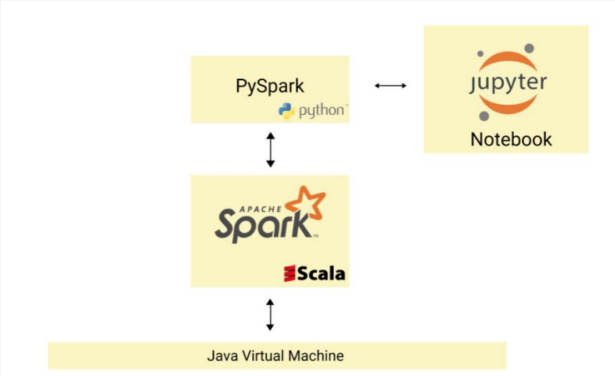
Faire du ML avec des données massives (Big Data)

Installation de spark et de Jupiter

Mode Local

l'application Spark s'exécute sur une seule machine. Ce mode sert à

construire et tester vos prototypes d'applications Spark sur votre propre machine



Mode Cluster

l'application Spark s'exécute sur plusieurs machines.

Ce mode cluster est utile lorsque vous voudrez exécuter votre application Spark sur plusieurs machines dans un environnement cloud comme Amazon Web Services ou Microsoft Azure.

Installer la JVM

Installer la JVM (c:\jdk) :

https://www.oracle.com/fr/java/technologies/javase/javase8-archive-downloads.html

Vérifier l'installation: java -version

Ajouter JDK au PATH (Variable d'environnement)

Spark fonctionne sur la machine virtuelle Java (JVM), qui est fournie avec le kit de développement Java SE (JDK). Nous allons installer la version JDK 8u202.

Installer Python

Installer la JVM (c:\jdk) :

https://www.python.org/downloads/

Vérifier l'installation: python --version

python™

Installer Spark

Installer Spark version précompilée :

http://spark.apache.org/downloads.html

c:\spark

Utilitaire de simulation de Hadoop:

https://sundog-spark.s3.amazonaws.com/winutils.exe

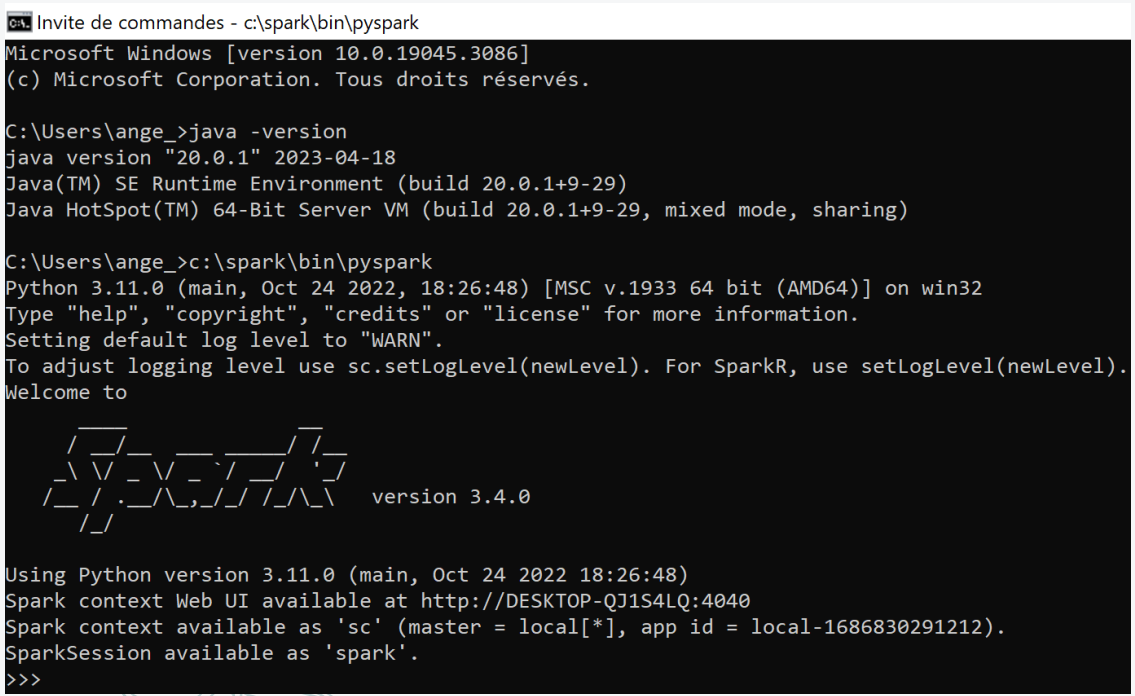
C:\winutils\bin\winutils.exe

Configuration des Variables d’environnement


Exécution de Pyspark
pyspark

Lancer pyspark

Ouvrir l'invite de commande et lancer la commande c:\spark\bin\pyspark.



Intégrer Jupyter notebook

Installer la bibliothèque Python findspark:

pip install findspark

Installer jupyter

pip install jupyter

Installer notebook

pip install notebook

PARTIE 2

Exploration de PySpark

Exploration de quelques méthodes de base PySpark

La structure de données centrale de Spark est le RDD (Resilient Distributed Dataset).

RDD ➔ Ensemble de données distribuées dans la RAM d'un cluster de plusieurs machine

Les transformations : map(), filter(), reduceByKey()

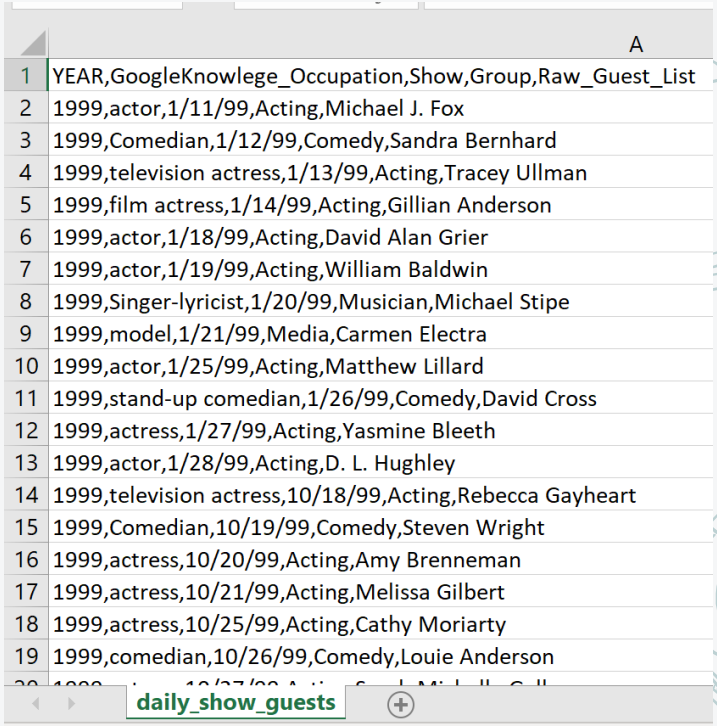
Les actions : take(), saveAsTextFile(), count(), collect()

Le jeu de données

Expérimentation des différentes méthodes avec le fichier daily_show_guest.csv

invités ayant déjà participés à l'émission américaine The DailyShow.

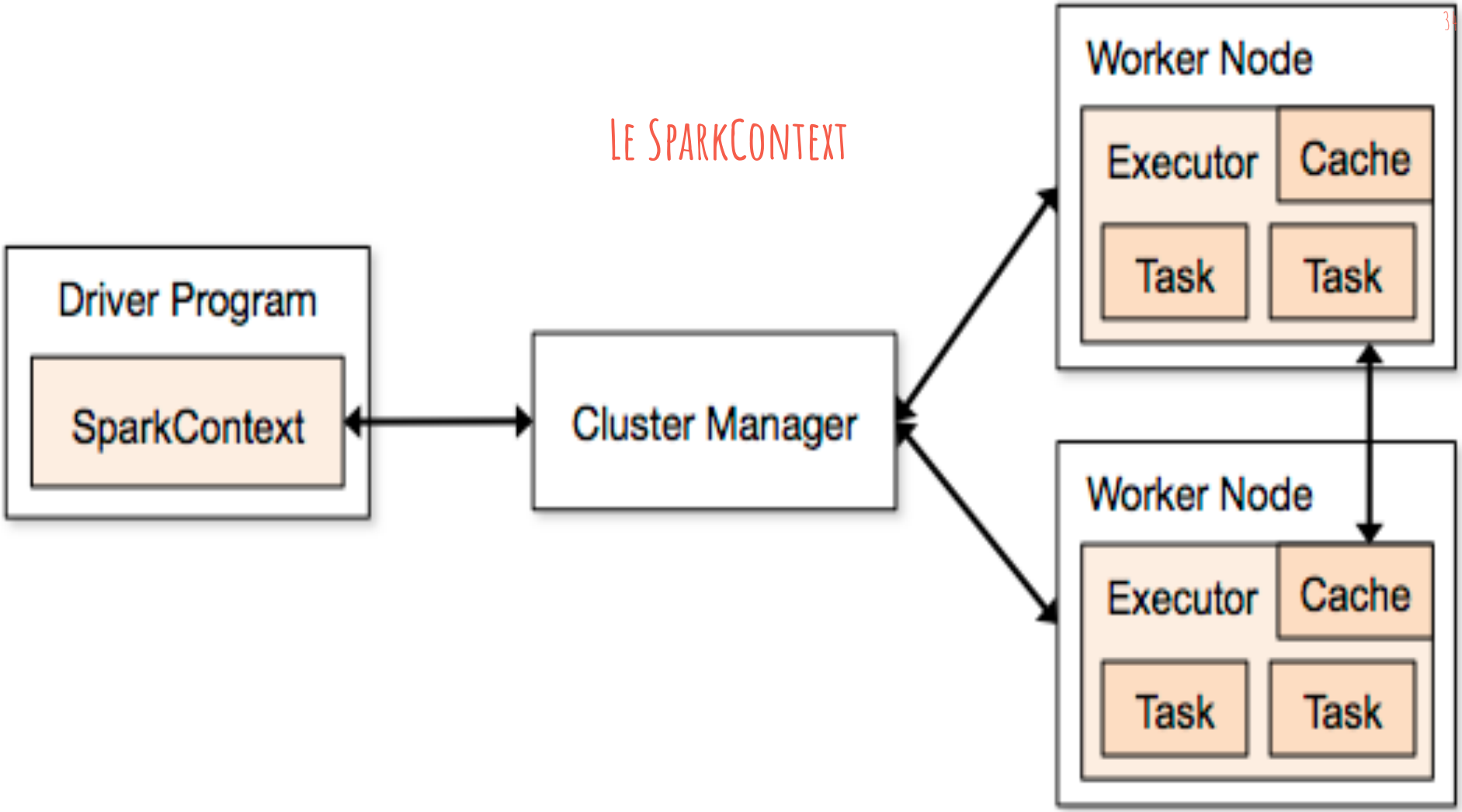
https://raw.githubusercontent.com/fivethirtyeight/data/master/daily-show-guests/daily_show_guests.csv



	A
1	YEAR,GoogleKnowlege_Occupation,Show,Group,Raw_Guest_List
2	1999,actor,1/11/99,Acting,Michael J. Fox
3	1999,Comedian,1/12/99,Comedy,Sandra Bernhard
4	1999,television actress,1/13/99,Acting,Tracey Ullman
5	1999,film actress,1/14/99,Acting,Gillian Anderson
6	1999,actor,1/18/99,Acting,David Alan Grier
7	1999,actor,1/19/99,Acting,William Baldwin
8	1999,Singer-lyricist,1/20/99,Musician,Michael Stipe
9	1999,model,1/21/99,Media,Carmen Electra
10	1999,actor,1/25/99,Acting,Matthew Lillard
11	1999,stand-up comedian,1/26/99,Comedy,David Cross
12	1999,actress,1/27/99,Acting,Yasmine Bleeth
13	1999,actor,1/28/99,Acting,D. L. Hughley
14	1999,television actress,10/18/99,Acting,Rebecca Gayheart
15	1999,Comedian,10/19/99,Comedy,Steven Wright
16	1999,actress,10/20/99,Acting,Amy Brenneman
17	1999,actress,10/21/99,Acting,Melissa Gilbert
18	1999,actress,10/25/99,Acting,Cathy Moriarty
19	1999,comedian,10/26/99,Comedy,Louie Anderson

daily_show_guests

Le SparkContext



La fonction Map()

La fonction **map(f)** applique une fonction f à chaque élément du RDD.

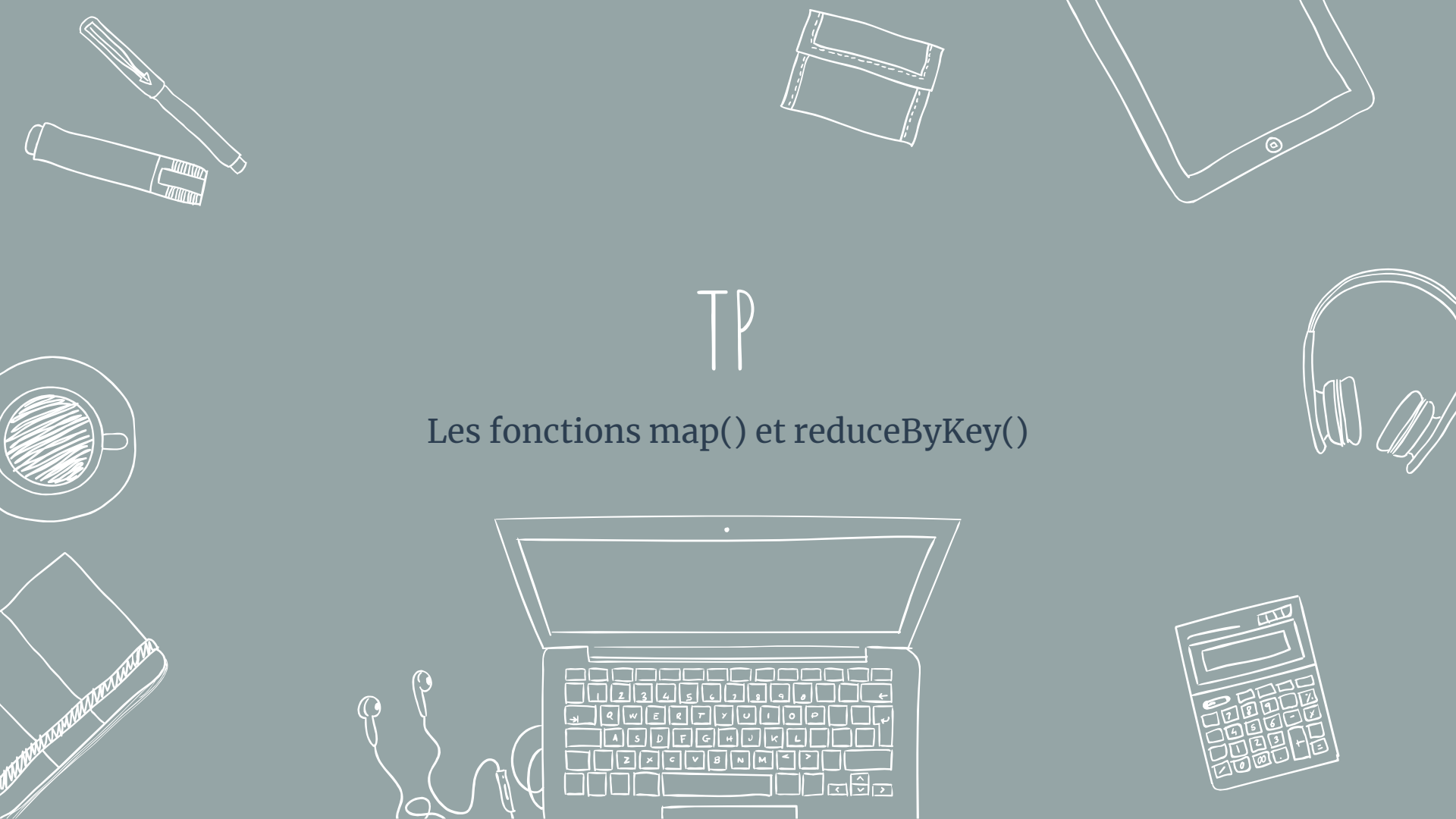
Comme les RDD sont des objets itérables (comme la plupart des objets Python), Spark exécute la fonction f à chaque itération et renvoie un nouveau RDD.

La fonction ReduceByKey()

La méthode **reduceByKey(f)** va combiner les tuples ayant des clés identiques en utilisant la fonction spécifiée entre parenthèses.

TP

Les fonctions map() et reduceByKey()



La fonction filter()

Spark dispose d'une fonction **filter(f)** qui crée un nouvel objet RDD en filtrant un objet RDD existant selon des critères spécifiques.

Si nous spécifions une fonction f qui renvoie une valeur binaire, **True** ou **False**, le RDD résultant sera composé d'éléments pour lesquels la fonction a été évaluée à **True**.

TP

The Daily Show

Merci!

Des Questions?

Vous pouvez me contacter:

Ange_amon@live.fr

+225 07 48 36 70 76