### Université Constantine2 Avril 2024

### Faculté NTIC, dépt. IFA M1 -SDIA

### Interrogation 01

**Nom:**

**Prénom:**

**Groupe:**

### (TDSAI)

### Exercice 1

Nous voulons extraire certaines informations sur le dataset ([personne.csv](https://drive.google.com/drive/folders/1bsnGcgs8oQMVuVQpgYJCvlwbp_DlUtc_dXT3dhQD4xfNvlws79OxwREYYLAUXIRPcxvRhu4u)) ces informations sont représentées par les colonnes suivantes (Nom, Prénom, Genre, Age, Poids, Taille).

from pyspark.sql import SparkSession

sc=SparkSession.builder.master('local[\*]').getOrCreate()

df = sc.read.csv('personne.csv',header=True)

Ecrire un code SparkSQL qui répond aux requêtes suivantes :

1. Afficher le genre de personnes qui sont moins de 85 kg.

df.createOrReplaceTempView('pers')

sc.sql('select Genre, Poids from pers where Poids<85').show()

df.filter(df.Poids<85).select(df['Genre'],df['Poids']).show()

1. Afficher les noms et prénoms des personnes qui sont plus de 1.65 m`

sc.sql('select Nom, Prenom, Taille from pers where Taille>1.65').show()

df.filter(df.Taille>1.65).select(df.Nom,df.Prenom,df.Taille).show()

1. Incrémenter l’âge de chaque personne par 2 ans puis afficher leurs noms et le nouvel âge.

sc.sql('select Nom, Age+2 from pers').show()

df.select(df['Nom'],df.Age+2).show()

### Questions : cochez la ou (les) bonne(s) réponse(s)

1. Dans Spark, l’opération textFile() est utilisée pour créer :

a) les RDDs à partir de fichier de type .txt uniquement b) les RDDs à partir de fichier de type .csv uniquement

c) les RDDs à partir de fichier peu importe son type d) les DataFrames

1. En Spark ‘inferSchema=True’ signifie :

a) parcourir un fichier .csv et mettre chaque colonne de type String

b) création d’un RDD c) création d’un DataFrame

d) parcourir un fichier .csv et déduire le schéma de chaque colonne

1. Les 4 modules de l’écosystème de Spark sont :

a) SparkSQL b) RDD c) Spark Streaming

d)Yarn e) MLib f) GraphX g) Mahout

**Bonne chance**