### Université Constantine2 Avril 2024

### Faculté NTIC, dépt. IFA M1 -SDIA

### Interrogation 01

**Nom:**

**Prénom:**

**Groupe:**

### (TDSAI)

### Exercice 1

Soit le jeu de données relatif aux ventes de matériel informatique, les données sont stockées dans un fichier « sales.csv » représentées par trois colonnes (compagnie, personne, ventes).

1. Que fait le code suivant ?

from pyspark.sql import SparkSession

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

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

**Réponse :**créer une session Spark puis crée un DataFrame ‘df’ à partir d’un fichier ‘sales.csv’.

1. Ecrire un code SparkSQL qui répond aux requêtes suivantes :
2. Afficher les ventes inferieurs à 1500 Da.

df.filter(df.sales<1500).show()

df.createOrReplaceTempView('ventes')

sc.sql('select \* from ventes where sales<1500').show()

1. Afficher le maximum des ventes par personne.

from pyspark.sql.functions import max

df.groupBy(df.person).agg(max('sales')).show()

sc.sql('select person, max(sales) from ventes GROUP BY person').show()

1. Afficher la moyenne des ventes de la compagnie ‘Intel’.

from pyspark.sql.functions import mean

df1=df.filter(df.company=='Intel').select(mean(df.sales)).show()

sc.sql('select mean(sales) from ventes where company =="Intel"').show()

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

**1.** Laquelle des structures suivantes est la structure de données fondamentale de Spark :

a) DataFrame b) RDD c) DataSet d) toutes les réponses

**2.** Quelles sont les opérations paresseuses de Spark ?

a) les actions b) les transformations c) aucune réponse d) les réponses a et b ensembles

**3.** Quelle méthode est utilisée pour récupérer tous les éléments d’un RDD ?

a) show() b) Coalesce() c) take() d) collect() e) lookup()

**Bonne chance**