ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

SUJET D'EXAMEN BLANC

MATIERE : Big Data et Data Mining

NIVEAU :	**3GESI**	**OPTION :**	**GESI**	**SEMESTRE :**	**S8**
SESSION :	**Normale**	**DUREE :**	**2H**	**CONSIGNES :**	**Documents Non autorisés**
ANNEE UNIVERSITAIRE 2025-2026		**ENSEIGNANT (E)**	**Abdelali EL Gourari**		
!!! TOUS LES EQUIPEMENTS SMART DOIVENT ETRE ETEINTS ET RANGES !!!					

EXERCICE 1 : Fondamentaux du Big Data (5 points)

Partie A - QCM et questions courtes (2,5 points)

1. [0,5 pt] Parmi les caractéristiques suivantes des données, laquelle correspond à la VARIÉTÉ dans les 5V du Big Data ?
    a) La vitesse de génération des données
    b) La diversité des formats et sources (texte, images, vidéos, JSON...)
    c) La fiabilité et la qualité des données
    d) La valeur commerciale extraite des données
2. [0,5 pt] Quelle est la différence fondamentale entre des données structurées et des données non structurées ? Donnez un exemple concret de chaque type dans le contexte d'une plateforme e-commerce.
3. [0,5 pt] Expliquez en quoi la VÉRACITÉ représente un défi majeur pour les entreprises exploitant le Big Data.
4. [0,5 pt] Dans le contexte des capteurs IoT (Internet des Objets), identifiez lesquels des 5V sont particulièrement pertinents et justifiez votre réponse.
5. [0,5 pt] Pourquoi les méthodes traditionnelles de stockage (bases de données relationnelles classiques) sont-elles insuffisantes face au Volume du Big Data ?

Partie B - Étude de cas (2,5 points)

Un hôpital moderne souhaite mettre en place une stratégie Big Data pour améliorer la qualité des soins. Les données concernées incluent :

- Dossiers médicaux électroniques (données patients, prescriptions, résultats d'analyses)
- Images médicales (radiographies, IRM, scanners - plusieurs téraoctets par an)

- Données des capteurs de surveillance (fréquence cardiaque, tension, température en temps réel)
- Publications scientifiques et articles de recherche
- Données de satisfaction des patients (commentaires, enquêtes)

Questions :

1. [1 pt] Pour chaque type de données ci-dessus, identifiez s'il s'agit de données structurées, semi-structurées ou non structurées. Justifiez chaque classification.
2. [1 pt] En vous appuyant sur les 5V, analysez les défis spécifiques auxquels l'hôpital sera confronté. Proposez une solution technique adaptée pour chaque défi identifié.
3. [0,5 pt] Quelle valeur (au sens de la 5ème V) l'hôpital peut-il espérer extraire de ces données ? Citez deux cas d'usage concrets.

EXERCICE 2 : Architecture et Écosystème Hadoop (5 points)

Partie A - Architecture HDFS (2,5 points)

Un cluster Hadoop est composé d'un Namenode et de 5 Datanodes. Le facteur de réplication est configuré à 3. Un fichier de 900 Mo doit être stocké. Les blocs HDFS ont une taille de 128 Mo.

1. [0,5 pt] Combien de blocs le fichier sera-t-il découpé ? Justifiez par le calcul.
2. [0,5 pt] Combien de copies de chaque bloc existeront-t-elles dans le cluster ? Quel est le nombre total de blocs stockés (toutes réplicas confondus) ?
3. [0,5 pt] Décrivez le rôle du Namenode dans ce processus de stockage. Quelles métadonnées conserve-t-il ?
4. [0,5 pt] Que se passe-t-il si le Datanode stockant le bloc B2 tombe en panne ? Décrivez le mécanisme de tolérance aux pannes d'HDFS.
5. [0,5 pt] Un client souhaite lire ce fichier. Décrivez les étapes du processus de lecture, en précisant les interactions entre Client, Namenode et Datanodes.

Partie B - Composants Hadoop et YARN (2,5 points)

1. Expliquez le rôle de YARN (Yet Another Resource Negotiator) dans l'architecture Hadoop. Pourquoi a-t-il été introduit ? Quels problèmes résout-il par rapport à Hadoop 1.x ?
2. [1 pt] Décrivez le cycle de vie d'une application MapReduce sous YARN : depuis la soumission du job jusqu'à la récupération des résultats. Identifiez les acteurs principaux (ResourceManager, NodeManager, ApplicationMaster) et leurs interactions.

EXERCICE 3 — Analyse de ventes avec Pandas

Une entreprise possède un fichier `ventes.csv` contenant les informations suivantes :

produit	categorie	prix	quantite	ville
PC Portable	Informatique	8000	3	Marrakech
Souris	Accessoires	150	10	Casablanca
Clavier	Accessoires	300	5	Rabat
Écran	Informatique	2500	2	Marrakech
Imprimante	Informatique	1800	1	Fès

```
import pandas as pd

# Chargement des données
df = pd.read_csv("ventes.csv")

# Affichage des premières lignes
print(df.head())

# Produits dont le prix est supérieur à 1000
result1 = df[df["prix"] > 1000]

# Création d'une nouvelle colonne
df["montant_total"] = df["prix"] * df["quantite"]

# Calcul du chiffre d'affaires total
ca_total = df["montant_total"].sum()

# Produits vendus à Marrakech
result2 = df[df["ville"] == "Marrakech"]

# Produit avec la quantité maximale
result3 = df[df["quantite"] == df["quantite"].max()]
```

1. Expliquer le rôle de l'instruction suivante : `df = pd.read_csv("ventes.csv")`(1 pt)

2. Que permet d'afficher la fonction head() ?(1 pt)

3. Expliquer la logique de cette instruction : `df[df["prix"] > 1000]`(2 pts)

4. Quelle est l'utilité de la colonne montant_total ?(2 pts)

5. Quelle opération réalise la fonction sum() dans cet exercice ?(1 pt)

6. Donner le résultat attendu de : `df[df["ville"] == "Marrakech"]`(2 pts)

7. Pourquoi Pandas est-il adapté à l'analyse de données en entreprise ?Donner deux avantages.(1 pt)

EXERCICE 4 — Nettoyage et analyse de données étudiants

Une école possède des données concernant ses étudiants.

```
import pandas as pd
import numpy as np

data = {
    "nom": ["Ali", "Sara", "Amal", "Yassine", "Salma"],
```

```
    "age": [21, 22, np.nan, 23, 20],
    "note": [14, 18, 12, np.nan, 16],
    "filiere": ["IA", "Big Data", "IA", "Cybersécurité", "Big Data"]
}

df = pd.DataFrame(data)

# Informations générales
print(df.info())

# Détection des valeurs manquantes
print(df.isnull())

# Remplacement des valeurs manquantes
df["age"] = df["age"].fillna(df["age"].mean())
df["note"] = df["note"].fillna(0)

# Moyenne des notes
moyenne = df["note"].mean()

# Étudiants avec note >= 15
result1 = df[df["note"] >= 15]

# Tri décroissant des notes
result2 = df.sort_values(by="note", ascending=False)

# Moyenne des notes par filière
result3 = df.groupby("filiere")["note"].mean()
```

1. Quel est le rôle de DataFrame dans Pandas ?(1 pt)

2. Pourquoi utilise-t-on isnull() ?(1 pt)

3. Expliquer le rôle de : `fillna()`(1 pt)

4. Pourquoi remplacer des valeurs manquantes avant l'analyse des données ?(2 pts)

5. Que représente : `df["note"].mean()`(1 pt)

6. Expliquer le résultat attendu de : `df[df["note"] >= 15]`(2 pts)

7. Quel est l'intérêt de : `groupby("filiere")`(1 pt)

8. Expliquer pourquoi le nettoyage des données est une étape importante en Big Data et Data Science.(1 pt)