

TP1: Programmation des RDDs avec Spark

Exercice 1 :

1. On souhaite développer une application Spark permettant, à partir d'un fichier texte (ventes.txt) en entrée, contenant les ventes d'une entreprise dans les différentes villes, de déterminer le total des ventes par ville. La structure du fichier ventes.txt est de la forme suivante :

date ville produit prix

Vous testez votre code en local avant de lancer le job sur le cluster.

2. Vous créez une deuxième application permettant de calculer le prix total des ventes des produits par ville et par année.

Exercice 2 : Analyse de fichiers de logs avec RDD en Java

On dispose d'un fichier de logs d'un serveur web au **format Apache**. Chaque ligne représente une requête envoyée par un client au serveur, contenant des informations comme l'adresse IP, la date, la ressource demandée, le code HTTP de la réponse, etc.

Exemple de format général :

```
IP - identifiant_utilisateur [date:heure +zone] "méthode URL protocole"  
code_HTTP taille "referer" "user-agent"
```

Exemple de fichier access.log :

```
127.0.0.1 - - [10/Oct/2025:09:15:32 +0000] "GET /index.html HTTP/1.1" 200  
1024 "http://example.com" "Mozilla/5.0"  
192.168.1.10 - john [10/Oct/2025:09:17:12 +0000] "POST /login HTTP/1.1" 302  
512 "-" "curl/7.68.0"  
203.0.113.5 - - [10/Oct/2025:09:19:01 +0000] "GET /docs/report.pdf  
HTTP/1.1" 404 64 "-" "Mozilla/5.0"  
198.51.100.7 - - [10/Oct/2025:09:25:48 +0000] "GET /api/data?id=123  
HTTP/1.1" 500 128 "-" "PostmanRuntime/7.26.8"
```

```
192.168.1.11 - jane [10/Oct/2025:09:30:05 +0000] "GET /dashboard HTTP/1.1"  
200 4096 "http://intranet" "Mozilla/5.0"
```

Travail demandé :

1. **Lecture des données :**

Charger le fichier de logs dans un RDD à partir du système de fichiers local.

2. **Extraction des champs :**

À partir de chaque ligne, extraire :

- L'adresse **IP** du client,
- La **date/heure**,
- La **méthode HTTP** (GET, POST, ...),
- La **ressource demandée** (ex. /index.html),
- Le **code HTTP** (200, 404, 500, ...),
- La **taille** de la réponse.

3. **Statistiques de base :**

- Le **nombre total de requêtes**,
- Le **nombre total d'erreurs** (codes HTTP ≥ 400),
- Le **pourcentage d'erreurs** par rapport au total.

4. **Top 5 des adresses IP** ayant effectué le plus de requêtes.

5. **Top 5 des ressources** les plus demandées.

6. **Répartition des requêtes par code HTTP** (200, 302, 404, 500, etc.).