TP 4: Spark SQL

Format du fichier CSV

Id, titre, description, service, date

Solution (PySpark – Spark SQL)

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import year

spark = SparkSession.builder.appName("IncidentsSQL").getOrCreate()

df = spark.read.option("header", True).option("inferSchema", True).csv("incidents.csv")

df.createOrReplaceTempView("incidents")
```

1. Nombre d’incidents par service

```
SELECT service, COUNT(*) AS nombre_incidents
FROM incidents
GROUP BY service;
```

2. Les deux années avec le plus d’incidents

```
SELECT YEAR(date) AS annee, COUNT(*) AS nombre_incidents
FROM incidents
GROUP BY YEAR(date)
ORDER BY nombre_incidents DESC
LIMIT 2;
```

Exécution des requêtes

```
spark.sql("""
SELECT service, COUNT(*) AS nombre_incidents
FROM incidents
GROUP BY service
""").show()

spark.sql("""
SELECT YEAR(date) AS annee, COUNT(*) AS nombre_incidents
FROM incidents
GROUP BY YEAR(date)
ORDER BY nombre_incidents DESC
LIMIT 2
""").show()
```