Ex. No. 2	**Word Count Using Apache Spark**
Youtube Link	**https://youtu.be/-szhD_VHMDg**
Date of Exercise	6.10.25

AIM

To implement a **Word Count** program using Apache Spark and analyze how distributed processing works with transformations and actions.

Procedure:

1. **Create Input File :**

   Apache Spark is fast

   Spark is open source

   Spark runs in memory

2. **Launch PySpark:**
3. **Read the input file:**
4. **Tokenize lines into words:**
5. **Map each word to a key-value pair:**
6. **Aggregate the counts by word:**
7. **Display the result**

Program:

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("WordCount").getOrCreate()
sc = spark.sparkContext

lines = sc.textFile("sample.txt")
words = lines.flatMap(lambda line: line.split())
wordPairs = words.map(lambda word: (word.lower(), 1))
wordCounts = wordPairs.reduceByKey(lambda a, b: a + b)
for word, count in wordCounts.collect():
    print(f"{word} {count}")

spark.stop()
```

Output:

```
hello 3
world 2
spark 2
big 3
data 3
processing 1
is 1
great 1
for 1
```

Result :

The Word Count application was successfully implemented using Apache Spark, demonstrating distributed processing using RDD transformations and actions.