In [1]:
```
from pyspark.sql import SparkSession

spark = (
    SparkSession
    .builder
    .master("local[*]")
    .appName("Spark Streaming")
    .getOrCreate()
)
```

In [2]:
```
spark
```

Out[2]:

SparkSession - in-memory

SparkContext

Spark UI

Version	v3.3.0
Master	local[*]
AppName	Spark Streaming

In [17]:
```
df_raw = spark.read.format("text").load("datas/input.txt")
df_raw.show()
```

```
+--------------------+
|               value|
+--------------------+
|this is some text...|
+--------------------+
```

In [18]:
```
from pyspark.sql.functions import split

df_words = df_raw.withColumn("words", split("value", " "))
df_words.show()
```

```
+--------------------+--------------------+
|               value|               words|
+--------------------+--------------------+
|this is some text...|[this, is, some, ...|
+--------------------+--------------------+
```

In [19]:
```
from pyspark.sql.functions import explode

df_explode = df_words.withColumn("word", explode("words"))
df_explode.show()
```

```
+--------------------+--------------------+--------+
|               value|               words|    word|
+--------------------+--------------------+--------+
|this is some text...|[this, is, some, ...|    this|
|this is some text...|[this, is, some, ...|      is|
|this is some text...|[this, is, some, ...|    some|
|this is some text...|[this, is, some, ...|    text|
|this is some text...|[this, is, some, ...|     for|
|this is some text...|[this, is, some, ...|   spark|
|this is some text...|[this, is, some, ...|      as|
|this is some text...|[this, is, some, ...|   input|
|this is some text...|[this, is, some, ...|      to|
|this is some text...|[this, is, some, ...| analyze|
|this is some text...|[this, is, some, ...|      it|
|this is some text...|[this, is, some, ...|     and|
|this is some text...|[this, is, some, ...|    does|
|this is some text...|[this, is, some, ...|    some|
|this is some text...|[this, is, some, ...| process|
|this is some text...|[this, is, some, ...|    like|
|this is some text...|[this, is, some, ...|counting|
|this is some text...|[this, is, some, ...|     the|
|this is some text...|[this, is, some, ...|   words|
|this is some text...|[this, is, some, ...|     and|
+--------------------+--------------------+--------+
only showing top 20 rows
```

In [20]:
```
df_explode = df_explode.drop("value", "words")
df_explode.show()
```

```
+--------+
|    word|
+--------+
|    this|
|      is|
|    some|
|    text|
|     for|
|   spark|
|      as|
|   input|
|      to|
| analyze|
|      it|
|     and|
|    does|
|    some|
| process|
|    like|
|counting|
|     the|
|   words|
|     and|
+--------+
only showing top 20 rows
```

In [24]:
```
from pyspark.sql.functions import count, lit

df_counted = df_explode.groupBy("word").count()
df_counted.show()
```

```
+--------+-----+
|    word|count|
+--------+-----+
|    some|    2|
|   input|    1|
| process|    1|
|     for|    1|
|counting|    1|
|   words|    1|
| attempt|    1|
|    with|    1|
|   other|    1|
|      is|    2|
|      it|    1|
|    does|    1|
|  things|    1|
|   spark|    2|
|    done|    1|
|    file|    1|
|     the|    1|
| analyze|    1|
|    like|    1|
|     and|    2|
+--------+-----+
only showing top 20 rows
```