# Master PySpark: From Zero to Big Data Hero!!

## String Functions

1. Convert the first letter of each word to uppercase (initcap):

```
df.select(initcap(col("Country"))).show()
```

▸ (3) Spark Jobs

```
+----------------+
|initcap(Country)|
+----------------+
|             Usa|
|           India|
|         Germany|
|       Australia|
|           Japan|
|          Brazil|
+----------------+
```

Note: This transforms the first letter of each word in the Country column to uppercase.

2. Convert all text to lowercase (lower):

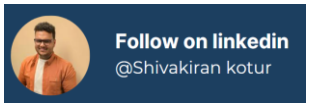
```
df.select(lower(col("Country"))).show()
```

▸ (3) Spark Jobs

```
+--------------+
|lower(Country)|
+--------------+
|           usa|
|         india|
|       germany|
|     australia|
|         japan|
|        brazil|
+--------------+
```

Note: Converts all letters in the Country column to lowercase.





3. Convert all text to uppercase (upper):

```
df.select(upper(col("Country"))).show()
```

▸ (3) Spark Jobs

```
+--------------+
|upper(Country)|
+--------------+
|           USA|
|         INDIA|
|       GERMANY|
|     AUSTRALIA|
|         JAPAN|
|        BRAZIL|
+--------------+
```

Note: Converts all letters in the Country column to uppercase.

Concatenation Functions

1. Concatenate two columns:

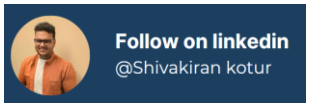
```
df.select(concat(col("Region"), col("Country"))).show()
```

▸ (3) Spark Jobs

```
+----------------------+
|concat(Region, Country)|
+----------------------+
|       North AmericaUSA|
|              AsiaIndia|
|          EuropeGermany|
|       OceaniaAustralia|
|              AsiaJapan|
|    South AmericaBrazil|
+----------------------+
```

Note: Concatenates the values of Region and Country without any separator.





2. Concatenate with a separator:

```
df.select(concat_ws(' | ', col("Region"), col("Country"))).show()
```

Note: Concatenates the values of Region and Country with | as a separator.

3. Create a new concatenated column:

```
concatenated_df = df.withColumn("concatenated", concat(df["Region"], lit(" "), df["Country"]))
concatenated_df.show()
```

(3) Spark Jobs

concatenated_df: pyspark.sql.dataframe.DataFrame = [Country: string, Region: string ... 3 more fields]

```
+---------+-------------+---------+---------+--------------------+
|  Country|       Region|UnitsSold|UnitPrice|        concatenated|
+---------+-------------+---------+---------+--------------------+
|      USA|North America|      100|     50.5|   North America USA|
|    India|         Asia|      300|     20.0|          Asia India|
|  Germany|       Europe|      200|     30.5|      Europe Germany|
|Australia|      Oceania|      150|     60.0|   Oceania Australia|
|    Japan|         Asia|      120|     45.0|          Asia Japan|
|   Brazil|South America|      180|     25.0|South America Brazil|
+---------+-------------+---------+---------+--------------------+
```

Note: This creates a new column concatenated by combining Region and Country with a space between them.

Summary of Key Functions:

- String Manipulation: You can convert strings to lowercase, uppercase, or capitalize the first letter of each word. Use initcap(), lower(), and upper() for these transformations.
- Concatenation: Use concat() to join two columns or concat_ws() to join with a separator.

These functions and transformations are common in PySpark for manipulating and querying data effectively!