LAB 7: PYSPARK using databricks

In [ ]:

```
pip install findspark
```

```
Python interpreter will be restarted.
Collecting findspark
  Using cached findspark-2.0.1-py2.py3-none-any.whl (4.4 kB)
Installing collected packages: findspark
Successfully installed findspark-2.0.1
Python interpreter will be restarted.
```

In [ ]:

```
import findspark
```

In [ ]:

```
findspark.init()
```

In [ ]:

```
from pyspark.sql import SparkSession
```

In [ ]:

```
Spark=SparkSession.builder.master("local").appName("Word count").getOrCreate()
```

In [ ]:

```
sc=Spark.sparkContext
```

In [ ]:

```
in_text=sc.textFile("dbfs:/FileStore/shared_uploads/ds215229142@bhc.edu.in/New_Text_Document.txt")
```

In [ ]:

```
in_text.collect()
```

```
Out[7]: ['Databricks SQL provides an easy-to-use platform for analysts who want to run SQL queries on their data la
ke,',
 'create multiple visualization types to explore query results from different perspectives, and build and share das
hboards']
```

In [ ]:

```
text_flat=in_text.flatMap(lambda x :x.split(' ')).map(lambda word:(word,1)).reduceByKey(lambda x,y:x+y)
```

In [ ]:

```
text_flat.collect()
```

```
Out[9]: [('SQL', 2),
 ('provides', 1),
 ('an', 1),
 ('platform', 1),
 ('run', 1),
 ('multiple', 1),
 ('query', 1),
 ('results', 1),
 ('different', 1),
 ('perspectives,', 1),
 ('share', 1),
 ('Databricks', 1),
 ('easy-to-use', 1),
 ('for', 1),
 ('analysts', 1),
 ('who', 1),
 ('want', 1),
 ('to', 2),
 ('queries', 1),
 ('on', 1),
 ('their', 1),
 ('data', 1),
 ('lake,', 1),
 ('create', 1),
 ('visualization', 1),
 ('types', 1),
 ('explore', 1),
 ('from', 1),
 ('and', 2),
 ('build', 1),
 ('dashboards', 1)]
```