Chapter 2: Overview of Spark

Ex 1: Cho tập tin t8.shakespeare.txt

- Cho biết tập tin có bao nhiêu line?
- Cho biết word "love" xuất hiện trên bao nhiêu line?
- Cho biết trong tập tin có xuất hiện số (bất kỳ) không? Nếu có thì xuất hiện trên bao nhiêu line? Liệt kê các line đó?

In [1]:
```
import findspark
findspark.init()
```

In [2]:
```
import pyspark
```

In [3]:
```
from pyspark import SparkContext
sc =SparkContext(master="local", appName="Ex1 Spark Context")
```

In [4]:
```
file_name = "t8.shakespeare.txt"
data = sc.textFile(file_name).cache()
```

In [5]:
```
print("Lines: %i" % (len(data.collect())))
```

```
Lines: 124456
```

In [6]:
```
numLoves = data.filter(lambda s: 'love' in s).count()
```

In [7]:
```
print("Lines with 'love': %i" % (numLoves))
```

```
Lines with 'love': 2484
```

In [8]:
```
import re
pattern = "[\s|\S]*\d{1,}[\s|\S]*"
```

In [9]:
```
numbers = data.filter(lambda s: re.match(pattern,s)).count()
```

In [10]:
```
if numbers > 0 :
    print("Lines with numbers: %i" % (numbers))
    number_lines = data.filter(lambda s: re.match(pattern,s))
    i = 0
    for line in number_lines.collect():
        print(i,":",line)
        i+=1
else:
    print("No number in this text.")
```

```
Lines with numbers: 1191
0 : This is the 100th Etext file presented by Project Gutenberg, and
1 : SHAKESPEARE IS COPYRIGHT 1990-1993 BY WORLD LIBRARY, INC., AND IS
2 : DISTRIBUTED SO LONG AS SUCH COPIES (1) ARE FOR YOUR OR OTHERS
3 : PERSONAL USE ONLY, AND (2) ARE NOT DISTRIBUTED OR USED
4 : **Etexts Readable By Both Humans and By Computers, Since 1971**
5 : January, 1994  [Etext #100]
6 : ******This file should be named shaks12.txt or shaks12.zip*****
7 : Corrected EDITIONS of our etexts get a new NUMBER, shaks13.txt
8 : VERSIONS based on separate sources get new LETTER, shaks10a.txt
9 : Please call them at 1-800-443-0238 or email julianc@netcom.com
10 : up to date first edition [xxxxx10x.xxx] please check file sizes
11 : per text is nominally estimated at one dollar, then we produce 2
12 : Files by the December 31, 2001.  [10,000 x 100,000,000=Trillion]
13 : which is 10% of the expected number of computer users by the end
14 : of the year 2001.
15 : P. O. Box  2782
16 : Champaign, IL 61825
17 : cd etext/etext91
```