BDA Lab 11

Your goal is to run Spark with a simple
application that will read a file, store its
content in a dataframe, and display the result.
You will learn how to set up your working
environment. You will also learn how to
interact with Spark and do basic operations.

Install Spark Locally on your Machine

Documentation @ https://pypi.org/project/pyspark/

Guide @ https://medium.com/tinghaochen/how-to-install-pyspark-locally-94501eefe421

1. Install Scala

2. Install Spark

3. Install Pyspark

4. Build a Spark context

5. Ingest books.csv by reading it into Spark Data Frame using:

```
session.read.csv(header=True, inferSchema=True, path=absolute_file_path)
```

6. Display first 5 rows

```
+---+--------+------------------+-----------+------------------+
| id|authorId|             title|releaseDate|              link|
+---+--------+------------------+-----------+------------------+
|  1|       1|Fantastic Beasts ...|  11/18/16|http://amzn.to/2k...|
|  2|       1|Harry Potter and ...|   10/6/15|http://amzn.to/2l...|
|  3|       1|The Tales of Beed...|   12/4/08|http://amzn.to/2k...|
|  4|       1|Harry Potter and ...|   10/4/16|http://amzn.to/2k...|
|  5|       2|Informix 12.10 on...|   4/23/17|http://amzn.to/2i...|
+---+--------+------------------+-----------+------------------+
only showing top 5 rows
```

Outcomes:

Although the example is simple, you would have accomplished the following:

- Installed all the components you need to work with Spark. (Yes, it is that easy!)
- Created a session where code can be executed.
- Loaded a CSV data file.

- Displayed five rows of this dataset.

Requirements:
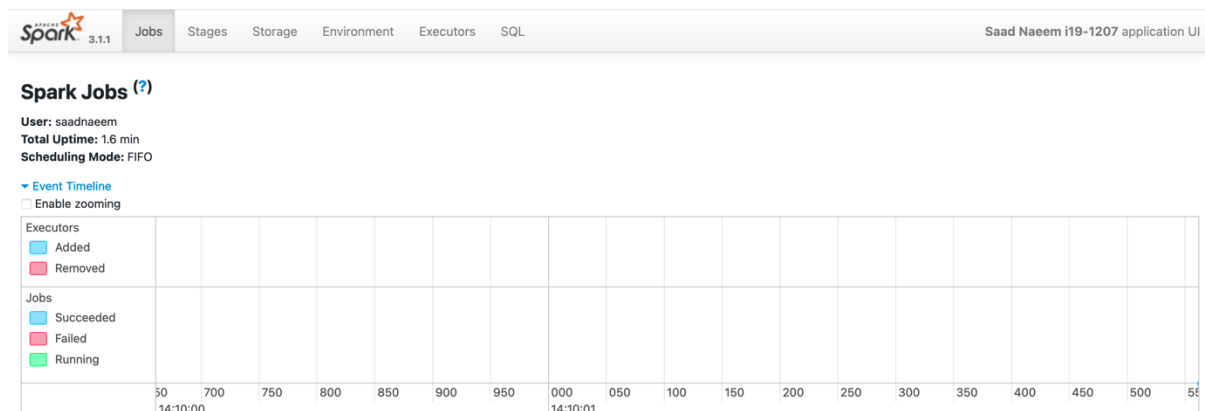
AppName would be your Name and Roll Number:

```
In [4]:  1  session

Out[4]:  SparkSession - in-memory
         SparkContext

         Spark UI
         Version
         v3.1.1
         Master
         local[*]
         AppName
         Saad Naeem i19-1207
```

Click on Spark UI to get the following screen shot that confirms the system user and the application name along with correct installation verification:

Finally, the Spark DataFrame showing the records:

```
In [6]:  1  # Shows at most 5 rows from the dataframe
         2  |

         +---+--------+------------------+-----------+--------------------+
         | id|authorId|             title|releaseDate|                link|
         +---+--------+------------------+-----------+--------------------+
         |  1|       1|Fantastic Beasts ...|   11/18/16|http://amzn.to/2k...|
         |  2|       1|Harry Potter and ...|    10/6/15|http://amzn.to/2l...|
         |  3|       1|The Tales of Beed...|    12/4/08|http://amzn.to/2k...|
         |  4|       1|Harry Potter and ...|    10/4/16|http://amzn.to/2k...|
         |  5|       2|Informix 12.10 on...|    4/23/17|http://amzn.to/2i...|
         +---+--------+------------------+-----------+--------------------+
         only showing top 5 rows
```

Do Not cut/leave out any part of the screenshot.