**February 27, 2024**

Creating a GitHub repository and testing Git Bash to push files from a local folder.

**February 28, 2024**

Starting reading about the assignment topic ***“combining Advanced data analytics (specifically Neural Networks) and Big Data Storage & Processing”.*** I’ve found interesting books to enrich my knowledge about this subject:

1. Mastering Apache Spark 2.x (Second Edition) by Romeo Kienzler. Packt.
2. Advanced Analytics with Spark by Sandy Ryza, Uri Laserson, Sean Owen and Josh Wills. O’Reilly.
3. Deep Learning with Python by Francois Chollet. Manning.
4. Large Scale Machine Learning with Spark by Rezaul Karim and Mahedi Kaysar. Packt.
5. Learning Spark by Holden Karau, Andy Konwinski, Patrick Wendell and Matei Zaharia. O’Reilly.
6. Mastering Machine Learning with Python in Six Steps, by Manohar Swamynathan. Apress.

I’ve installed Apache Spark 3.5.1 on my Windows 10 source:  
<https://www.youtube.com/watch?v=0F4fokX5MPQ>

To open Apache on http://desktop-ik25uha:4040: C:\Program Files\spark-3.5.1-bin-hadoop3\bin>spark-shell

I had to install Java Development Kit aka JDK. When you install Apache Spark or any other Java-based software, you need to have JDK installed on your system to ensure that the necessary tools and libraries are available for development and execution. Source:  
<https://www.youtube.com/watch?v=-O4QVijnA7Y>

The next step is to push records into Apache Spark and link a Jupyter Notebook to it, stablish the connection and pull records into a pandas dataframe.

Videos to connect Jupyter Notebooks to Apache:

Part 1:

https://www.youtube.com/watch?v=XvbEADU0IPU

Part 2:

<https://www.youtube.com/watch?v=e_QoFQjZwqc>

**February 29, 2024**

Researching about how push data into Apache Spark and connect it Jupyter Notebook.

Getting books suggested by Muhammad regarding Map reduce design patterns:

* Tom White, 2012, Hadoop The Definitive Guide, O’Reilly Publishing
* Hadoop with Python, Zach Radtka; Donald Miner, O'Reilly Media, Inc., 2015.
* Lublinsky B., Smith K. T. and Yakubovich A 2013, Professional Hadoop Solutions, Wrox [ISBN: 13:978-11186]
* Holmes A 2012, Hadoop in Practice, Manning Publications [ISBN: 13:978-16172]
* McKinney W. 2012, Python for Data Analysis, O'Reilly Media [ISBN: 13: 978-14493]

**March 1, 2024**

Connecting Jupyter Notebooks to Spark:

* Spark\_App\_Using\_PySpark.ipynb: Reading CSV file, creating database and table in Spark, printing all records in parket format, note that size decreases.
* PySpark\_Connection.ipynb: Simple app to see how we can interact with spark using pyspark.

**March 2, 2024**

When converting a Spark DataFrame to a Pandas DataFrame, it was not possible due to memory constraints. This is an interesting topic to discuss in the paper because many data scientists, when trying to develop machine learning models, will need to parse large Spark DataFrames into a Pandas DataFrame in order to perform machine learning or deep learning techniques. A nice takeaway for today.

**March 3, 2024**

Creating a Spark database from a large 3.61GB JSON file for modeling purposes is impractical due to its size. We need to adapt our dataset for modelling, below Jupyter Notebooks will demonstrate the entire process:

1. Creating\_Spark\_DB\_Loading\_3.61GB\_json.ipynb
2. Spark\_DB\_Parket\_pd\_df.ipynb

The error that made enlightened me to find an interesting topic:

**Py4JJavaError**: An error occurred while calling o56.collectToPython.

: org.apache.spark.SparkException: Job aborted due to stage failure: Total size of serialized results of 11 tasks (1145.7 MiB) is bigger than spark.driver.maxResultSize (1024.0 MiB)