**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)

**March 4, 2024**

Spark is often used for distributed data processing, including data cleaning, feature extraction, and handling large datasets that don't fit into a single machine's memory.

Keras, on the other hand, is a high-level neural networks API, typically used for building and training deep learning models.

Finding an example of a deep learning application using a dataset that could be classified under the "Big Data" umbrella, due to the large volume dataset.

Book:  
Deep Learning Convergence to Big Data Analytics.

**March 5, 2024**

Deep Learning, to demonstrate memory constraints on my laptop I am going to find a small dataset, pulled into MySQL, pulled using PySpark and then modelled it in an ANN.

Finally to demonstrate how Big Data has helped we will use a coding library (the likes of Keras) and implement some ANN. From the technical standpoint I believe this is a good approach to demonstrate how Big Data Storage/Processing and Advanced Data Analytics relate.

**March 6, 2024**

Importing 1.3GB csv file into MySQL using Jupyter Notebook. The most efficient way I’ve found to perform this task. It was impossible to accomplish using MySQL features.

Source: https://www.datablist.com/learn/csv/download-sample-csv-files

**March 7, 2024**

Researching articles:

1. Mobile Big Data Analytics Using Deep Learning and Apache Spark
2. Apache Spark and Deep Learning Models for High-Performance Network Intrusion Detection Using CSE-CIC-IDS2018

**March 8, 2024**

Article: A review of machine learning and big data applications in addressing

ecosystem service research gaps

**March 9, 2024**

<https://www.cloudduggu.com/spark/spark-sql/>

**March 10, 2024**

Geyer, R.C., Klein, T., and Nabi, M. (2017). 'Differentially Private Federated Learning: A Client Level Perspective', arXiv preprint arXiv:1712.07557.

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., & Yu, H. (2020). Federated Learning. Morgan & Claypool Publishers.

**March 11, 2024**

Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., and Bacon, D. (2016). 'Federated Learning: Strategies for Improving Communication Efficiency', *arXiv preprint arXiv:1610.05492.*

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. (2021). 'Federated Multi-Task Learning Under a Mixture of Distributions', *SIAM Journal on Optimization, 31(4), pp. 3075-3107.*

**March 12, 2024**

Title: Data Analytics and Storage Solutions: MySQL, Spark and APIs in Neural Network Contexts.

**March 13, 2024**

Timing pushing data into MySQL and Spark separately to evaluate performance.

**March 14, 2024**

Starting the paper.  
After today’s Big Data Storage class, Muhammad told that the approach of using Windows is not good as Hadoop can’t run here. To meet the learning outcome we must use Hadoop a Big Data storage like it.

**March 15, 2024**

Giving structure to the paper and changing paper title.

**March 16, 2024**

Completing topic Overview, objectives, research question, and starting the state of the Art.

**March 17, 2024**

Advancing State of Art section, word count 989. Leaving key papers and Literature review coming to the end for a more accurate opinion.

**March 18, 2024**

Assembling a neural network for the people.csv dataset, we aim for the model to predict gender based on job title and age. This is not an easy task; the loss is high, and the accuracy is low, making it challenging to develop a robust model that accurately predicts gender based on job title and age.

**March 19, 2024**

Writing structure for implementation part, uploading NN\_implementation jupyter Notebook used in VM Ubuntu.

**March 20, 2024**

Implementation Hadoop part completed, also dataset origin is mentioned and how size was increased. Uploading 1.Increasing\_dataset\_size.ipynb. Starting with MySQL dataload.

**March 21, 2024**

Uploading 2.Importing\_1.6GB\_CSV\_to\_MySQL.ipynb and 3. 3.HDFS\_Data\_to\_model\_NN.ipynb. First script is used to push records into MySQL and the second one is used to extract Hadoop data and model a NN.

**March 22, 2024**

Paper section ***D. Data from HDFS Modeled with a Neural Network*** completed. Next ***E. Data from MySQL Modeled with a Neural Network.***

**March 23, 2024**

Implementing NN using MySQL data *4.MySQL\_Data\_to\_model\_NN.ipynb*, we clearly see that Hadoop and Spark perform way better than MySQL RDBMS.

**March 24, 2024**

Creating script *5.API\_Data\_to\_model\_NN.ipynb*, and testing it.

**March 25, 2024**

*5.API\_Data\_to\_model\_NN.ipynb*, further improving and modelling the NN.

**March 26, 2024**

Implementation completed.

**March 27, 2024**

Starting the literature review, once the code is completed it is easier to link all the articles and sources.

**March 28, 2024**

Summarizing articles and integrating citations from all papers into a thematic literature review, which will encompass 20 articles.