**INFO-I 535 MANAGEMENT ACCESS USING BIG DATA**

**COURSE REFLECTION**

**Best Week:**

**1. Lifecycle & Pipelines Module:**

I selected this module based on my interest and will be helpful for my career.

**What specific actions or behaviors contributed to your performance that week?**

**Ans:** The effectiveness of my performance that week was significantly enhanced by engaging deeply with the module on creating data pipelines. I was particularly intrigued by the scalability of the ETL pipeline, which ensures that the system can handle increasing data volumes without performance degradation. This feature is crucial as it allows for the continuous integration of new data sources and growing data sets. Additionally, the ability of the pipeline to provide insights from the data is extremely beneficial. It supports data-driven decision-making by transforming raw data into actionable information, thus optimizing operational strategies, and improving business outcomes.

**How did you apply your study habits, time management, and learning strategies that week?**

**Study habits:** Dedicated specific times each day to learning, ensuring that each session had a clear objective related to understanding different aspects of learning and researching the topics about the white paper provided for that week. Applied concepts learned in tutorials by experimenting with creating mini-data pipelines using sample datasets. This took me the most time I had to understand resources like Virtual private cloud networks, how to configure that, how each step is performed in data pipeline. I couldn’t completely work on the filtering part.

**Time Management:** Prioritized learning modules based on their complexity and relevance, starting with basic concepts of data pipelines before moving to more advanced topics such as ETL processes and automation with tools like Airflow. This way I made sure that I wasn’t getting exhausted or procrastinating my work.

**Learning Strategies:** Started with the fundamentals of data pipelines and progressively advanced to more complex topics, ensuring a solid understanding before moving on. Participated in study groups to discuss challenging topics and share insights, which reinforced learning and provided different perspectives on the subject. This also helped me to verify that I was in the right track.

**Is there anything you could to improve that week (even if it was your best week)?**

While that week was highly productive, there was still room for improvement, particularly in my practical experience with configuration steps. I couldn't build a complete pipeline due to my lack of experience in configuring the necessary components effectively. If I had been more proficient in these steps, I would have been able to create a complete pipeline and visualize its output. This would not only have enriched my learning experience but also saved a considerable amount of time on my final project. Going forward, dedicating more time to understanding and practicing these configuration processes will be crucial. I constantly going wrong with creating the pipelines. I should work on clarifying with someone and then make sure my approach is on the right track.

**2. Processing and Analytics Module:**

I selected this module based on my interest and will be helpful for my career which aligns closely in Data Engineering field.

**What specific actions or behaviors contributed to your performance that week?**

For the above week, I focused intensely on the deep learning module, dedicating specific hours to study and ensuring I practiced the concepts through hands-on exercises. (Reading white paper, relevant YouTube videos, and article in medium). I actively participated in discussions, which helped reinforce my understanding, and what was my peers understanding and frequently revisited complex topics until I was confident with them.

**How did you apply your study habits, time management, and learning strategies that week?**

**Study habits:** Begin with reviewing the core concepts of Spark and PySpark, such as RDDs, transformations, and actions, then progress to applying these concepts in practical tasks. This includes setting up the Docker container for the PySpark environment, creating and handling RDDs, and performing transformations and actions on data. The most interesting part was understanding the importance of RDDs and their operations before moving on to more complex operations like MapReduce.

**Time Management:** Configuring the environment such as setting up the Docker container and with the Jupyter notebook interface. Next would be researching about RDD operations, and so forth. With the help of collaborative learning, I was able to ensure that operations are correct and used according to the assignment.

**Learning Strategies:** Involved myself by testing out commands, debugging issues, and modifying examples to see different outcomes this way understanding how each line is utilized.

**Is there anything you could to improve that week (even if it was your best week)?**

To further enhance my project where I visualized the top 10 frequent words with spam detection using PySpark basically building a text classification model. A brief idea that I would have done began by collecting and preprocessed a labeled dataset of spam and non-spam emails or messages after that utilize PySpark's MLlib to train a classification model, such as logistic regression or a decision tree, on your dataset or used both and compared their accuracy and precision.

**Worst (Didn’t find interesting) modules:**

**1. Data Modeling**

I found it less interesting compared to other modules because I felt it lacked practical insights.

**What specific actions or behaviors contributed to your performance that week?**

my performance was significantly bolstered by carefully structuring my study time to delve into each NoSQL data model, dedicating sessions to understanding their architecture and application. I consistently evaluated and adjusted my learning pace and focus areas based on my daily progress in grasping the complexities of Big Data challenges. Engaging in discussions with my peers and applying theoretical knowledge to practical datasets were key behaviors that enriched my learning experience and solidified my understanding of the material.

**How did you apply your study habits, time management, and learning strategies that week?**

**Study Habits:** In that week I extracted and broke it into pieces of information of the paper by breaking down NoSQL models and reinforced my learning by creating summaries and mind maps, especially for the six NoSQL data models reviewed, ensuring I understood their unique attributes and limitations.

**Time Management:** I allocated specific time blocks for each Big Data model, which allowed me to dive deep into their structures and uses without feeling overwhelmed this was interesting to read. Daily reviews of my progress were crucial, as they helped me adjust my schedule to spend more time on areas like the real-time data processing capabilities of each model. I ensured to intersperse breaks between sessions, maintaining a balance to prevent burnout and retain a high level of focus throughout the week.

**Learning Strategies:** I utilized a variety of learning resources, from academic papers to technical blogs, to see how these data models work in different scenarios, which broadened my comprehension. Collaborative discussions with peers about the 10 rules for scalable datastores provided diverse insights and enhanced my understanding of the material's practical relevance. I couldn’t implement anything practically because was working NoSQL for my other assignment.

**Is there anything you could to improve that week (even if it was your best week)?**

Even though it was a productive week, I have regrets or anything that can be done better. It is because I implemented cloud storage for a delivery management system.

**2. Impact of big data**

During that week, my performance was affected primarily by not dedicating enough time to thoroughly review and understand the course modules and related papers. This lack of preparation hindered my ability to actively engage in discussions. My study habits were not optimally utilized, as I didn't manage my time efficiently to accommodate comprehensive reading and comprehension of the materials provided. To enhance future performance, even if it were already a successful week, I could enhance my engagement by writing an article for a forum. This would not only help consolidate my understanding of AI fairness in critical sectors like banking but also allow me to share my insights more broadly, thereby reinforcing my learning and contributing to my academic and professional growth.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Submitted by Kirthivasan PN (kirpandu@iu.edu)