2023 Fall Session 6: Python Coding for SAS Programmers

1. **Q - Can we share the underlying reasons for Spark Pandas DataFrame being less performant than Spark DataFrame?**
   A - A Spark DataFrame is always faster than a Spark Panda DataFrame. That is because even though it is utilizing the Spark distribution on the compute, it still utilizes some of the Pandas functionality. Pandas was derived from the NumPy package and how it takes information and looks at it from a multiple dimension array, a matrix. There are certain functionalities and features that Pandas rely on the data being distributed in a certain way. Spark Pandas does distribute the data, but it does not distribute it as efficiently as a true Spark DataFrame distributes it.

2. **Q - Is there anything equivalent to SAS' dictionary tables in Python to find details about multiple tables in a schema?**
   A - The column object, col_count= len(hsp.info.columns) within the DataFrame can be utilized to determine what the names of the columns are. You can look at what different columns there are within the DataFrame itself. If you wanted to print them, you would use print(hsp_info.coumns). It is already put into this list functionality that allows you to do a lot of for operations onto it. To look at specific list of fields that you wanted to pass within the list it could easily utilize that column attribute fields =[‘transfer_out’, ‘adm’ ‘discharge’ ‘readm’ ‘unplanned_readms’] as well as a dtype attribute. You do not have to do this manually. This can automatically come up with a list of every single string or object column within your DataFrame. There is a dtype functionality as well. There is also the ability to extract the actual data type of all the columns within your DataFrame and then filter only those columns, much like you would within your SAS dictionary.

3. **Q - More of a general Databricks question, but is there functionality that writes something akin to a SAS log to Simple Storage Service (S3)? Does the functionality of sending emails within code exist in Databricks?**
   A - Yes to both. You can set up these notebooks to run as jobs. There are ways to externally print out the log in the workspace to like you would Proc print on the SAS side where you are externally printing your log and your output to your workbench. You can send emails and make application programming interface (API) calls in Databricks.

4. **Q - I have noticed that most PySpark functions are just wrappers for SQL code. Have you seen cases when there is not an SQL equivalent?**
   A - There has not been an equivalent noted in the CCSQ environment yet. A lot of things done within PySpark can be done within Spark SQL. There is no reason why you cannot do the exact same filtering and creating new fields based on case when statements on the SQL side.

5. **Q - Thanks for showing lag functionality for readmissions, would love to get code on how to roll continuing claims (status code = 30) into one claim in Python.**
   A - We would be happy to work with you and your team to look at the specifics of what

you are trying to achieve and get it working on the Python side. You can fill out a ServiceNow ticket to get started.