2024 Spring Session 2: Code Conversion Best Practices

Date: Wednesday, 5/15 2:00-3:00 pm ET
Speaker: Derek Cruikshank

1. **Q- Does this mean that we will have to combine R and Python within the same code?**
   A- You can use multiple languages in the same Databricks notebook, but in different cells.

2. **Q- Related to Grant's presentation about DataComPy and Spark, I have had trouble installing Spark Compare. Any idea why?**
   A- This may be a permissions issue in Databricks preventing installation that would require a #help-service-center-sos ticket submission to allow user to install packages. Full documentation on package available here: https://capitalone.github.io/datacompy/ Note: Update post-data camp has slightly changed some of syntax of the package and now uses datacompy.legacy for direct spark implementation.

3. **Q- Will we run into memory issues trying to run models on large data sets in R (outside of Spark)? For example, if a model type is not implemented in Spark.**
   A- In the future, we will be opening up Databricks jobs to allow you to run jobs outside of your specific notebook cluster, choosing different cluster sizes. More to come soon as we are working out the limits and capabilities.

4. **Q- How much memory do we have access to on a single cluster when running R or using Pa ndas?**
   A- If you are not utilizing Spark packages that allow you to distribute the processing, then you will be only utilizing the driver node which is 32GB in size. There is some cut off the top for overhead. However, in this program increment (PI) we will also be enabling Databricks jobs which will give you the ability to run jobs outside of the normal interactive notebook cluster and choose different size options.

5. **Q- Can the R cluster now access data sets in our schemas?**
   A- Not currently, but it is our priority in this PI.

6. **Q- For this common code repository, could we eventually develop them as functions and ha ve a package we could all import into our notebooks?**
   A- That will be available in the future.

7. **Q- In which Slack channel is all this going to be discussed?**
   A- Topic specific channels will be created, and we will make an announcement when they are available.

8. **Q- The ability to run jobs in Databricks has been available for a while. I wanted to confirm if you are just going to change some settings to allow users to select additional clusters for those jobs (which was not available before). If so, will users be able to freely expand the memory size for each job's cluster.**

A- The jobs that was refered to are actual job clusters that users will be able to utilize individually and separate from your org/group specific compute cluster. Like a personal compute.

9. **Q- Where did you save the data you got?**
   A- Data should be stored at the Simple Storage Service (S3) workbench.

10. **Q- We've had projects from CMS where we are not allowed to use ChatGPT due to privacy concerns. Is there guidance for when we can use ChatGPT to translate code that is approved by CMS ?**
    A- The guidance has been that there is absolutely no data sharing and none of the code should include Protected Health Information (PHI)/Personally Identifiable Information (PII)
    nor display any hostnames/servers.

11. **Q- What is the size limit for each file that can be saved on to the S3 workbench that Derek mentioned? Also, could you share code showing how to save the DataFrame to S3 workbench?**
    A- You would not share the DataFrame, as that is in-memory on the compute cluster. You can write the DataFrame out to a Centralized Data Repository (CDR) database table, or you can write it out as a text delimited type of file to your S3 workbench .

12. **Q- Do parquet and/or Python pickle tables count as text delimited for S3? I have done some basic reading, and those formats will allow us to keep some formatting for things like dates in our data that other for mats like comma separated values (CSVs) do not have.**
    A- Parquet is an especially useful and efficient file type to use and will be better than a text delimited file. In fact, most of the database tables in the CDR are made up of underlying parquet files.