2024 Spring Session 5: R for SAS Programmers

Date: Wednesday, 5/15 4:15-5:00 pm ET
Speaker: Daniel Anderson

1. **Q- What packages are available for me to use in Databricks?**
   A- Anything from the Comprehensive R Archive Network (CRAN) repository can be installed.

2. **Q- Do you have any troubleshooting tips?**
   A- For more on troubleshooting tips, see our Spring 2024 Data Camp Session 2.

3. **Q- When do you recommend R vs. Python or SQL?**
   A- R is for robust statical analysis. For more on R, Python and SQL, see our Spring 2024 Data Camp Session 2.

4. **Q- Will there be additional training for R?**
   A- Yes. There will be additional training during office hour sessions. Dates and times will be announced soon.

5. **Q- How can I read or write a SAS7BDAT file?**
   A- There are several packages available through the CRAN repository. E.g., Haven.

6. **Q- When using an R notebook, does dbfs:/tmp have any access controls or can everyone on the platform see data moved there?**
   A- If you are mounting a Simple Storage Service (S3) bucket to the Databricks File System (DBFS) using dbutils.fs.mount, then access is controlled the same way that it is with Access Management (IAM) roles and policies and S3 bucket policies for the specific group. Do not copy or move any data directly onto the DBFS because that is available to all users to see and access. Use the mount option, or just copy the files onto the local filesystem which is on the driver node.

7. **Q- Can Tibbles work with SparkR? Or does it have to be Sparklyr?**
   A- Tibbles only work with the Tidyverse ecosystem associated with the dplyr package.

8. **Q- Is there a reason you would need to use Spark(ly)R for reading in data? Is there that much processing involved that you would not just use a non-Spark version?**
   A- It depends on the size of the data in which you are reading. A non-Spark version is limited to only the available memory and central processing unit (CPU) on the single driver node compared to Spark having distributive processing across worker nodes and giving more available memory and CPU. Usually if you want to exceed the size of your worker node, then you will need to distribute your processing to multiple workers and need Spark.

9. **Q- I have noticed that the R cluster requires the dbfs prefix, but not the Python cluster (that does not work there). Can you expand on this difference between the clusters? It seems like they are both referring to the same physical space but would be good to know if that is not the case. Is there is a deeper underlying difference between how the files are accessed between R and Python clusters?**

   A- Because of some nuances in how Databricks does access and some remaining vestiges of the old system, R clusters cannot access the Centralized Data Repository (CDR) data directly and can only access the files/folders in workbenches. Python clusters have their metadata store directly tied to the CDR data and require less syntax than grabbing the files vs interfacing with the tables. In this Program Increment (PI) we are working to enable R clusters to work like the Python ones.

10. **Q- Do SparkR or SparklyR support data table?**
    A- Yes. Data Dot Table is part of the base structure. However, all your distributed computing work must be converted into a Spark DataFrame for it to work.