2023 Fall Session 4: Introduction to R Coding in Databricks

1. **Q - Does the R Databricks cluster automatically choose a mirror?**
   A - Yes and you can see the default mirror by running getOption("repos").

2. **Q - Can you use packages like SQLDF to still write SQL queries on the R only cluster?**
   A - You will not be able to run SQL commands on R-only clusters at this moment due to inability to instantiate Hive client on R clusters. This is a known issue and one that will change because of the Ambari decommission and metadata migration to Amazon Web Services (AWS) Glue. Instead, it would be better to run SQL commands on your organization's other cluster (ending in "_compute_cluster") either using Python or SQL language directly by invoking SQL in cell with "%sql".

3. **Q - Within Databricks, do you need to reinstall packages within each new notebook or are they saved somewhere (i.e., they would be on your PC)?**
   A - When you install a package within a notebook, it is only available for that notebook. When the cluster auto terminates due to no use, the installed package would get wiped out and need to be re-installed when you spin your cluster back up and re-open the notebook. If there's a specific package that you want permanently installed and available for your organization's cluster, you can reach out and request that from the Data & Analytics team.

4. **Q - Since the platform logs you out after X amount of time, what should I do if my statistical models take 8-10 hours to run?**
   A - Make sure that you are connected to your organization's r_only_compute_cluster. Also, running code as a Job under the Workflows section allows a background run even if you log out.

5. **Q - Why does it take approximately 30 minutes (i.e., more than a couple of minutes like it does on my local machine) to load packages (probably because it must first install the packages and then load but this process still is faster on my local machine)?**
   A - This can be due to a lot of different issues. Here are some considerations:

   - Cluster initialization - the Databricks R-only cluster first needs to be spun up, which can take several minutes to load. Make sure the cluster is active (green circle in top right of notebook) before running a cell to install a package.
   - Network latency - this could be due to your internet connection and the accessibility to external resources like Comprehensive R Archive Network (CRAN).
   - Package dependencies - depending on the package, there can be several dependencies that already might be saved to your local machine's R library and thus will not need to be re-installed. In the R cluster, if these dependent packages are not default

installations, they will also need to be installed. This can be seen from the output logs generated after install.packages() function is called.

If this is still an issue, you can put in a ServiceNow ticket for assistance.

6. **Q - Why do some functions work one day and stop working another day? system.file() is one of two examples that used to work but stopped working a week later (incident happened late October).**
   A - This seems like an issue that may be due to running a different language in a cell rather than R. system.file is a base R function and therefore should be available when running as an R command on an R cluster.

7. **Q - Can a large file be mounted to a memory mapped file and if so, how?**
   A - It depends on how large the file is. The specifications are the primary node for the R cluster for each organization. It also depends how much memory you have available within just that one primary node. It is 32 gigs and eight courses for the driver node. You can see how much free memory you have by running system("free -m") in R.

8. **Q - Clarification for the question above: can we freely use heap memory?**
   A - In Databricks clusters, especially in the context of Apache Spark, memory management is typically managed differently than in traditional R or other single-node environments. Databricks clusters use a distributed memory model, and the concept of "heap memory" is not directly applicable in the same way it is in a single-node R environment. Because of this, if reaching the limit of free memory for single-node use, it would benefit the user to avoid using heap memory altogether and restructure the notebook to use spark (or SparkR in the context of R). This way the notebook can efficiently store and process the data.

9. **Q - Can we have documentation in the Knowledge Base for how to mount and import/export data in both R and Python?**
   A - We are currently in the process of working through best practices documentation around this specific functionality. We are still working through some of the capabilities and security with the Databricks team, but we will have code snippets and documentation around how to create a mount with your Amazon Simple Storage Service (AWS S3) workbench bucket.

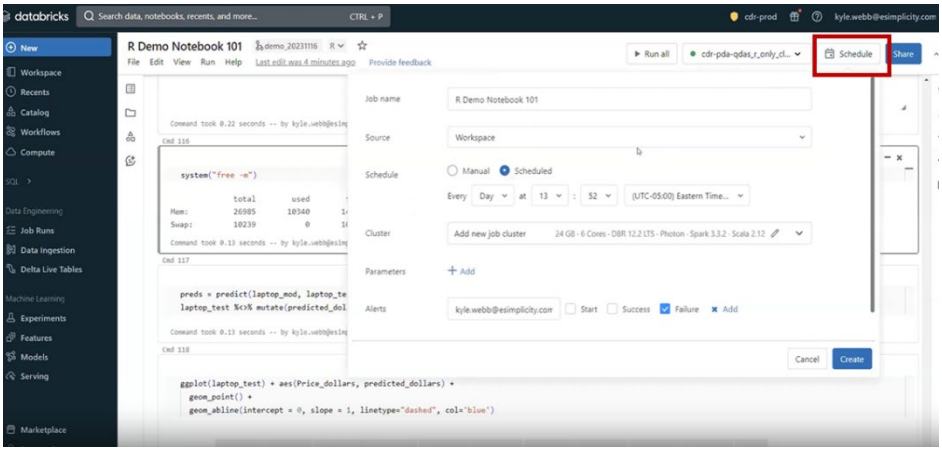
10. **Q - Are Centralized Data Repository (CDR) datasets already mounted?**
    A - No, they are not. CDR data is currently not available using the R-only compute cluster due to security issues. Once Ambari is decommissioned and we are migrated away from the Hive metadata catalog, the R cluster will be able to have access to the CDR tables like the Python compute cluster has.

11. **Q - Can you please share the information on how to get this notebook, if it is available now?**
    A – The notebook is under Session 4 on the Fall 2023 Data Camp page.

12. **Q - Is there a timeline on when the Internet Quality Improvement & Evaluation System (iQIES)-CASPER schema will be in Databricks? Now they are only available in memory in SAS in a CASLIB.**
    A - All iQIES data currently available through SAS CASLIBs will eventually be available from Databricks. Currently, there is no timeline, but it will be in the future.

13. **Q - Can you read a SAS Dataset into R?**
    A - There are ways to integrate SAS with R on the Python side. It is required that the sas7bdat Python package is used. It does require the Databricks File System (DBFS) mount functionality to then copy from the mount to the local file system for it to be able to be read. We are working on documentation regarding best practices on what we would recommend and how to utilize that functionality. As we work more with the Databricks team and the security team to ensure that everything is running smooth and what the subtle nuances are between file types and needing to keep it to a local file system versus being able to write directly onto the DBFS mount. There will be more to come. Another option is to convert the SAS dataset to a .csv table within SAS, then import the .csv file.

14. **Q - Does R (or the R notebooks) have the capability to schedule jobs to run at a future time similar to Viya?**
    A - Yes. You can schedule a job by clicking Schedule in the upper-right hand corner of the screen and fill out the information in the Schedule window.



15. **Q - Does knitting in Databricks work the same as R markdown in RStudio? Meaning you can knit to HTML, PDF, or Word directly from Databricks?**
    A - It does not. The easiest way to replicate the output from RStudio is to export the notebook as an RMarkdown or HTML file.