2023 Data Camp CCSQ D&A Introduction into Databricks

1. **When Databricks (DBX) is connected to SAS Viya will that have any impact on how we write structured query language (SQL) queries? For instance, right now we have hive_sql macros. Will we need to change these?**
A - The Data & Analytics team is working on creating Global macros that connects with Databricks and help on macros will be published on how to rewrite the queries if any.

2. **In my Databricks, I see a message that R is not there. It says that "Your administrator has only allowed Python and SQL commands on this cluster..." I was trying to run R. It is not clear to me how this decision was made and who to contact regarding any Databricks issues. I am looking to "fail fast" but it is not happening.**
A - The Cluster you are running on only allows python and sql. We have created another cluster with R enabled. Choose the R enabled cluster for running R code.

3. **I remember seeing Databricks as an option in my HCQIS Access Roles and Profile (HARP) request roles page. I no longer see it. Is there any change ?**
A - Complete the following HARP instructions:
1. Open QualityNet Analytics
2. Click the contractor name
3. Databricks DeveloperAll SAS Viya, Zeppelin, Databricks are at the same place.

4. **Some R packages need Devtools. Can we install Devtools on the R cluster?**
A - Currently users have all capabilities to install notebook scoped libraries. If there is a requirement for the libraries to be installed cluster-wide, a request must be made so that the admin team can install them at cluster level.

5. **Some R packages need RcppArmadillo, a special R package. Can we install RcppArmadillo on the R cluster?**
A - Currently users have all capabilities to install notebook scoped libraries. If there is a requirement for the libraries to be installed cluster-wide, a request must be made so that the admin team can install them at cluster level.

6. **Can we install our own R packages, developed by ourselves?**
A - Currently users have all capabilities to install notebook scoped libraries. If there is a requirement for the libraries to be installed cluster-wide, a request must be made so that the admin team can install them at cluster level.

7. **How to install R packages, e.g., install into our own library or is there a common library shared by all users? If the latter, how do I control the versions of the R packages installed by different users? Will there be documentation on how to best approach this?**
A - Currently users have all capabilities to install notebook scoped libraries. If there is a requirement for the libraries to be installed cluster-wide, a request must be made so that the admin team can install them at cluster level.

8. **During the early stages of the roll out, there were issues with creating tables in DBX and then using those within SAS Viya via HiveQL code. Has that been resolved?**

   A - The resolution is currently being built out and tested. The issue will be resolved before release.

9. **When more DBX users come on board, will we see a drop in compute space performance ?**
A - Databricks is configured for the users to auto-scale on their org clusters. Clusters are created keeping the usage in mind. If there is a specific need to increase beyond the allotted, Decision Analysis & Resolution (DAR) request must be raised.

10. **Can I access the iQIES-CASPER-SCHEMA in Databricks? I do not believe it is in the Centralized Data Reposition (CDR). The way I access it in SAS is with: PROC CASUTIL; LOAD INCASLIB="iQIES-CASPER-SCHEMA"**
A -CASLIBs integration is in progress but Internet Quality Improvement and Evaluation System (iQIES) is slated for later.

11. **I am the Quality Improvement and Evaluation System (QIES)/ Automated Survey Process Environment (ASPEN) Coordinator for Georgia and am not sure if I need access to Databricks?... ASPEN users? What data is retrieved via Databricks?**

    A - QIES users do not have access to request Databricks currently. Access through HARP for iQIES groups and subgroups are different and onboarding is not ready for these groups.

12. **Is Tableau available now?**
A - No, Tableau is a QualityNet service tool. QuickSight is our preferred tool equivalent.

13. **It will be cool if the DBX platform can automatically see your prog code and route to the appropriate cluster rather than have user actively select it.**
A - We can put a request to Databricks on this question. To detect the language in the notebook and assign the R cluster or Python/SQL cluster but Databricks does not have the ability right now.

14. **Is parallel processing across multiple Central Processing Units (CPUs) allowed in the R environment?**
A - This will need specific packages to be installed cluster-wide. We have not tested this yet.

15. **As a HARP user, I am able to open QuickSight and see shared dashboards. What do I need to do to gain access to creating my own dashboards?**

    A - You would likely need an expanded role, like a deployer or developer. Here is the link to role information for QuickSight: Requesting & Getting Access

16. **Is there some plan to allow users access to the DBFS (Databricks File System)? or will there be some process to expedite mounting of Simple Storage Service (S3) buckets to Databricks? I am gathering metadata and metrics from data sets that would be very useful to have stored for deeper analysis.**

    A - DBFS is not enabled as it does not have anti-virus (AV) capabilities. To expedite mounting S3 buckets to Databricks users must fill out a DAR request.

17. **Do we need to request this role now to continue using SAS Viya?**

    A - No, you do not need a Databricks role to continue using Viya.

18. **Can we expand on the 60-minute limit? What are the options if they do end up running longer?**

    A - Sixty minutes limit is if there are no processes running, it will auto-terminate.

    The cluster spins down after 60 minutes of inactivity. If you are running something on the cluster that takes longer than 60 minutes, the cluster would not shut down because it is active throughout that entire time.

19. **If the clusters time out in 60 minutes, do the analyses run that quickly?**

A - Yes, the analysis does run much quicker than our other tools which is a huge benefit of the service. To clarify, inactive clusters terminate after 60 minutes, not if they are running analysis, but very few analyses take that long.

20. **If we schedule a job, do we need to prestart the compute cluster?**

    A - Jobs can be executed by scheduling through job clusters which is different from compute or R cluster.

    Currently, users have the capability to spin job clusters.

21. **When will R get access to Hive?**

    A - We are working on making this access available soon.

22. **You mentioned BYOA regarding the SQL updates to DBX for the June release earlier. What is BYOA in this case?**

    A - BYOA stands for Bring Your Own Analytics.

23. **What does this mean for the community? Do they bring their own solutions/tools into the CMS environment? What is this process?**

    A - Users bring their own analytics tools. The BYOA process is still under development right now, but requests right now should be routed through our request form.

24. **Is SAS Viya going away sometime soon? If yes when?**

    A - SAS Viya will be available for the near future, DBX will be the SQL warehouse and engine that will be below SAS Viya.

25. **Why are the CDR and Integrated Data Repository (IDR) both running as similar data systems but with some differences? Are there plans to combine them and keep the upsides of either/both?**

    A - The CDR is CCSQ's analytics data storage solution and includes program data that the IDR does not. The IDR is meant for enterprise users and does not include the CCSQ program data. Schemas for the same dataset may also be different between the two (like claims).

26. **Are there limited organizations that allow access to QuickSight? My organization, Great Plains QIN, is not on the list to select a new role to create dashboards.**

    A - CCSQ QuickSight is open to use by anyone in the CCSQ Community with a HARP ID. If a shared folder does not exist for a contract group, one can be requested. Details on how to request a shared folder can be found on the Requesting and Getting Access page in Confluence.

27. **Is the connection between SAS Viya and DBX now available?**

    A - That is Phase 2. More communications will be sent out shortly on this and when exactly it will be released.

28. **Will SAS Viya be available along with Databricks or do we have to switch to Databricks sometime in future? Will Databricks be the only platform option for the next statement of work (SOW) for analytics?**

    A - SAS Viya will be available for the near future, DBX will be the SQL warehouse and engine that will be below SAS Viya.

29. **Any way that CMS will re-run CMS Medicare Severity Diagnosis Related Groups (MS-DRGs) grouper upon request from user on a subset of data? Just wondering if accessing the grouper is something on the radar.**

    A - Yes. Users can request this via the Data and Analytics Request Form on Confluence. In the form, users should include their business case and justification.

30. **We have tested the collaboration mechanism within Databricks and found that across two teams, we used the same compute server. Is that expected and is that okay?**

    A - It is possible that the users could be in multiple groups and could be seeing both the compute clusters.