CCSQ Data & Analytics: Notebook Alternative Session

1. **Any view about R in Databricks? R package availability? Statistical models and such?**
   A - Yes, Databricks has several ways to support R. Here's a good overview: https://docs.databricks.com/sparkr/index.html. Aside from being able to run SparkR within a cluster, it supports running R normally as you would in a single-machine configuration. In this scenario, Databricks can be used to manage server resources so that they can be appropriately sized for your small, medium, and large R jobs. Users will also have the ability to install R packages on their own.
2. **Does either tool connect to workbenches?**
   A - We will have support within both tools to be able to read data that's stored in SAS Workbenches.
3. **What about query performance comparison to current Zeppelin and HIve?**
   A - This was not evaluated because of the multi-tenancy configuration in CDR (Zeppelin + Hive) today (recall that all CCSQ users access one single large CDR cluster). We are not able to control the CDR environment to isolate queries and evaluate performance for comparison.
4. **What does the transition to these tools look like?**
   A - As we productionalize the tool and platform that is selected by CMS, we will share our migration strategies in future communications call, which will include setting up accounts, providing guidance and best practices for your code, and work with teams directly.
5. **For the comparison of Databricks and Sagemaker, was the configuration of data access comparable? For example, was Databricks accessing in-memory data and, if so, was this also true for Sagemaker?**
   A - The method to accessing data sets were the same on both platforms as the underlying data sets were stored in S3 buckets
6. **Can Databricks connect to CDR to allow users to query claims?**
   A - If Databricks is selected as the compute platform to move forward with, it will replace CDR. Users will do their Python/R/Scala notebook development in Databricks notebooks, which will connect to Databricks compute clusters (e.g. their own "Personal CDR"). SAS Programs will connect to Databricks compute clusters for querying instead of the legacy CDR clusters.
7. **If we didn't participate in the pilot, how can we access Databricks and start using it for coding in R or Python?**
   A - Ken will touch on this at the roadmap 4pm EST session. The Pilot is currently closed and we are moving toward a selection. If data brick is chosen as the Ambari/Zeppelin replacement it will connect to the CDR based claims replacement
8. **What is a cluster for?**
   A - Notebooks will connect to compute clusters to run the code. In the Databricks scenario, users will have access to their own "Personal CDR" compute cluster that will scale with their analytics and would not conflict with jobs running from other groups.
9. **Can json files generated out of zeppelin be imported into DB?**
   A - Yes, JSON files can be loaded and read into Databricks for further analysis.
10. **Will users be able to install any Python/R packages or will this require pre-approval process first?**
    A - Python/R package installation will be available as a self-serve capability and scoped to your notebook so that they do not interfere and conflict with others on your team. Some select libraries are more involved, however, and can be installed with support from the D&A team
11. **Will integration with local IDEs be supported, such as connecting Databricks to VS Code and editing/running code from there?**
    A - You will not need a local IDE as Databricks notebooks provides that for you for your code development
12. **I found the logging and troubleshooting within Zeppelin to be suboptimal....does DB provide more refined outputs?**
    A - Logs (stdout, stderr, log4j) are present on the cluster UI in real-time and also available historically to help triage issues. Additionally, users can access the SparkUI for performance troubleshooting on clusters, and SQL queries on the Databricks SQL product have a full query profiler tool available.
13. **When spinning up a cluster in Databricks, it presents you an option "auto terminate after x hours" - Is it recommended to turn this option on or off ? Do we have to turn the cluster on manually everytime we run a notebook ?**
    A - This is a cost-saving feature that CMS can optionally use. Cold cluster starts typically take about 5 minutes.
14. **If the data are stored in Hadoop clusters, how will they be transferred to data bricks clusters?**
    A - Any/all data sets that are created in the current CDR is going to be accessible in Databricks once you are migrated over. Recall the efforts we spoke about in previous communications call that related to HDFS Migration efforts to S3. That work was to prepare for this evolution to our NextGen CDR Platform.