Take Home Programming Assessment

Instructions

This take home programming exercise involves converting the provided legacy R code to PySpark. Please review the R code and translate its functionality into Python using the PySpark and/or Pandas API on Spark. Ensure that your PySpark implementation is efficient, readable, and follows best practices for distributed computing. Where possible, avoid or minimize the use of user-defined functions. This assessment is meant to mimic a 5 point user story and should be completed in 2-3 days. Perform any validation of your converted code base as necessary and include this in your final solution packet.

Start with the start_script.R and change the necessary import paths to the location in which the uncompressed folder is located on your machine. The necessary packages for convertMe and the helper functions in helper.r will be installed and imported by this script. The script will also execute convertMe using the input data sources so there is an example of the expected output to which you can compare the output of your Pyspark code.

Databricks Design Considerations

Please discuss how you would modify the design of the code to leverage Databricks features. Consider the following:

- How would you utilize Databricks clusters for efficient processing?
- How would you take advantage of Databricks notebooks for collaboration and documentation?
- How would you incorporate Databricks Delta Lake for data storage and management?
- How would you leverage Databricks jobs for scheduling and automation?
- What other Databricks features would you use to simplify the given code base?