WiseWithData

SPROCKET

SPROCKET White Paper Series #2

Top 50 questions to ask vendors about migrating from SAS to PySpark or Databricks

WiseWithData

Ignite Open Analytics

There are number of vendors claiming to be able to convert SAS code to other languages, but few have any real experience converting it at scale. The following is a list of key questions you may want to ask your vendor, to ensure you get the facts on how they would handle performing a SAS migration to PySpark or Databricks.

General Information and Approach

1. Will you be converting the code manually or using automation? If automation, please provide more details (developed in-house or purchased product).
2. Please provide the number of people who will have access to the code along with locations.
3. How long would it take your team to convert 10,000 lines of Base SAS ETL code? How many people would be required for that time-frame?
4. Does your code use PySpark best practices (i.e. is it DataFrame based instead of RDD/ Python UDF based)? When do you use RDD's or UDF's?
5. Do you provide a commitment to code accuracy, or take an SLA approach to code delivery?
6. Is your solution a big bang approach, or is it flexible allowing us to work on code conversion and out task more challenging and critical processes to you?
7. Are there any additional libraries or API's required/recommended? If so, can you describe their function and installation process.
8. Does any of your converted code use the Python Pandas library? If so, why?
9. What DevOps practices do you leverage in delivering converted code? What industry standard tools do you use to enable those practices?
10. What versions of PySpark (or Databricks) are supported/recommended?

Support

11. What versions of SAS code do you support?
12. Do you support reading SAS datasets (sas7bdat files) in your generated code? If so, how?
13. How do you recommend we educate our SAS programmers to utilize a new programming language?
14. Do you provide a service to assist with development of a migration plan and prioritization of the appropriate processes?
15. How does your organization deal with project changes, such as additional processes to convert, changing direction to work with specific users or groups, or changes to the environment and where the data resides?
16. Is your service capable of adapting to meet the needs of our individual teams and users? Is there a process to identify and prioritize accordingly?
17. What kind of support is provided post code migration/conversion?
18. If the converted code is not performing properly in production, what steps would you take to resolve the issue?
19. Do you have provide PySpark support services (education, infrastructure, coding support)? Do you provide any free community support materials? If so, please provide details and experience.
20. Do you have experience migrating code from on On-Premise to Cloud based services (Full Cloud or Hybrid Cloud)? If so, what is your approach to helping you customers through the migration? What challenges have you encountered?

SAS Feature Coverage

21. Does your automation handle Proc SQL queries that are incompatible with SparkSQL (i.e. calculated columns, re-merging of summary statistics, etc.)? Is so, please describe your approach.
22. How do you handle built-in and custom formats and informats? Are they supported in datasteps and other procedures (SQL, Means, etc.)? Does your solution scale?
23. How many functions and call routines do you support? How many of them use DataFrame or SQL operations vs UDF's? Are they all supported in both datasteps and proc sql?
24. Which functions and call routines do you *not* support in automation and why?
25. Do you support Macro custom functions, macro logic and built-in macro functions? If so, what is your approach?
26. How do you support SAS Libraries and SAS/ACCESS engines? Do you support implicit and explicit pass-thru in proc SQL?
27. How do you handle datastep code that contains retain, by, or "first." and "last." statements?
28. How do you handle macro variables and their use?
29. Do you support "call symput" and PROC SQL's "into" statements? If so, please provide details on how?
30. How do you handle missing values? What are the challenges in supporting SAS missing values?

Accuracy and Consistency

31. What is your internal QA process? Do you have internal code standards?
32. Is your generated code PEP-8 compliant? Are there exceptions, if so for what reasons?
33. Do you keep comments, blank lines and overall code flow? In what circumstances do you generate code where logic is out of order with the original code?
34. What steps, or services are utilized to provide validation ensuring the outputs from SAS are equal to the outputs from PySpark (UAT)? Do you provide validation tools for this purpose?
35. What steps do you take in your generated code to handle null safety issues?
36. What sas functions do you support that do not have exact matches in the generated PySpark code and why do they not match?
37. Do you have a standard mapping of SAS data types to PySpark? What issues might be encountered between the data type remapping?
38. Under what circumstances are differences in validation are to be expected?
39. What SAS system options can affect output data results of SAS code? How do you handle those issues?
40. Do you have a formal bug reporting process to avoid repeating the same issues in the future?

Experience

41. Have you experienced any challenges with SAS to PySpark code conversion? If so, what did you do to overcome the issue?
42. What types of SAS code can cause performance problems when converted into PySpark?
43. Have you ever uncovered bugs in Python or PySpark? If so how did you resolve or work around them?
44. What problems might occur when converting a datastep merge statement into PySpark? How do you solve for those issues?
45. What's the largest single process you've converted? How many lines was it?
46. Have you migrated code generated within Enterprise Guide? How does that process work?
47. Have you migrated automatically generated code from DI Studio?
48. Have you migrated Enterprise Miner scoring model code? If so, what model types do you have experience with? What are the challenges in migrating this code?
49. In what circumstances would you re-engineer a code segment while converting it to PySpark?
50. Beyond base SAS, what SAS other products do you have experience converting? What challenges have you encountered in converting code from those products?