**PySpark Overview**

1. How to Create Partition by month and year in pyspark.
2. How to read variable no of columns in pyspark.
3. How to skip first few rows in pyspark data frame.
4. How to remove duplicate rows in pyspark Data frame.
5. How to read all files from nested folder in pyspark data frame.
6. How to get no of rows at each partition in pyspark data frame.
7. How to get no of rows from each file in pyspark data frame.
8. How to add sequence generated surrogate key as a column in data frame.
9. How to get column wise null records count in pyspark.
10. Why should not use crc32 fro key generation.
11. How to handle double delimiter and multi-delimiter in pyspark.
12. How to handle multiprocessing in Azure Data bricks.
13. How to take table ddl backup in Azure Data bricks.
14. Convert pyspark string to date
15. How to handle duplicate table in error in delta table.
16. How to handle bad data in pyspark.
17. How to get 53 week number in pypsark.
18. How to handle complex json file in pyspark.
19. Difference between orderby ,sort and
20. Difference between coalesce and repartition in pysaprk.
21. Pyspark word count program.
22. Dynamically processing complex json files.
23. How do I select a column name with spaces in pyspark.

Date Format in taditional db systems :

dd-mm-yyyy

Date Format in Pyspark:

yyyy-MM-dd

--> Remove duplicate Data

Can be done by four types

Distinct()

dropDuplicates()

Window function

Groupby with