30 PySpark Coding Question

Karthik Kondpak

Swipe for more

Question 1 Find the top N most frequent words in a large text file

Link: https://www.linkedin.com/posts/karthik-kondpak_pyspark-day-1-activity-7238755222674644992-20gD?utm_source=share&utm_medium=member_desktop

Question 2

Calculate the average salary and count of employees for each department.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday2-ugcPost-7238844784386138113-554k?utm_source=share&utm_medium=member_desktop

Question 3
Different ways to remove duplicates
Write a PySpark code to remove duplicate rows based on specific columns.

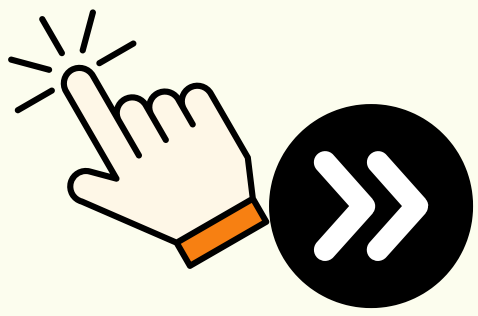
Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday3-activity-7239491612651745280-WKxK?utm_source=share&utm_medium=member_desktop

Question 4
Given a DataFrame df with columns id, name, and salary, write a PySpark code to filter rows where salary is greater than 5000 and select only the name column.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-4-activity-7239837651837579266-P5qm?utm_source=share&utm_medium=member_desktop

Question 5
How would you handle null values in a DataFrame? For example, drop rows with null values in the age column.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday5-activity-7240207575579770881-7UdJ?utm_source=share&utm_medium=member_desktop



Swipe for more

https://www.seekhobigdata.com/

Question 6

Adding a New Column to a DataFrame

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-6-activity-7240562417745674240-mT53?utm_source=share&utm_medium=member_desktop

Question 7

You are given two DataFrames in PySpark:
employee_df: Contains employee information.
department_df: Contains department information.
You need to perform an inner join on these DataFrames to find out which department each employee belongs to.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-7-ugcPost-7242027671709229056-WXH2?utm_source=share&utm_medium=member_desktop

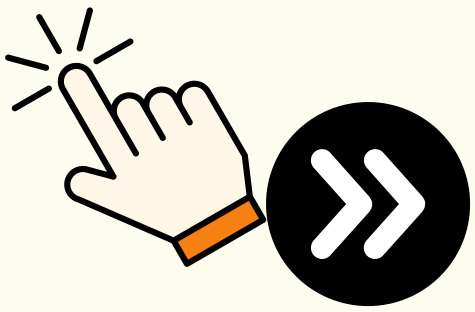
Question 8

You are given two DataFrames: employees: Contains employee details with the columns: emp_id, name, and dept_id.
1. departments: Contains department details with the columns: dept_id and dept_name.
2. You need to perform a left join on employees with departments to get all employee details, including the department name.
If an employee doesn't have a department, their department name should be null.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday8-activity-7242378122203914240-HSCT?utm_source=share&utm_medium=member_desktop

Swipe for more

https://www.seekhobigdata.com/



Question 9

You are working as a Data Engineer for an e-commerce company. The company has two datasets:
Orders Dataset: Contains order details such as order_id, customer_id, and order_status.
Customers Dataset: Contains information about customers like customer_id, customer_name, and customer_city.
Your task is to generate a report that contains all customer information, even if they have not placed any orders. Use a right join to solve this problem,
so that we get all customers, including those without orders.

Link : https://www.linkedin.com/posts/karthik-kondpak_pyspark-practice-day-9-ugcPost-7242748129903575040-8P3T?utm_source=share&utm_medium=member_desktop

Question 10

You are given a dataset containing daily stock prices. Write a PySpark program to calculate the running total
of stock prices for each stock symbol in the dataset.

Link : https://www.linkedin.com/posts/karthik-kondpak_pyspark-practiceday-10-activity-7243223763142508544-ntZE?utm_source=share&utm_medium=member_desktop

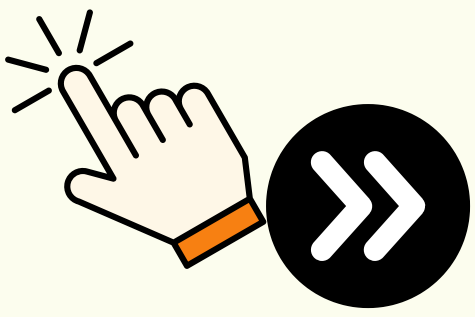
Question 11

In PySpark, there are multiple ways to rename columns in a DataFrame. Below are a few methods to achieve this, each explained with sample data and code:

Link : https://www.linkedin.com/posts/karthik-kondpak_pyspark-practice-day-11-activity-7243471796648214528-uTtq?utm_source=share&utm_medium=member_desktop

Swipe for more

https://www.seekhobigdata.com/



Question 12
How you can create a new column derived from existing columns in a PySpark DataFrame in different ways.

Link : https://www.linkedin.com/posts/karthik-kondpak_pyspark-practice-day-12-activity-7243830184762621952-BpOT?utm_source=share&utm_medium=member_desktop

Question 13
Sort a DataFrame based on one or multiple columns in PySpark.

Link : https://www.linkedin.com/posts/karthik-kondpak_pyspark-practice-day-13-activity-7244208662750117889-jfDu?utm_source=share&utm_medium=member_desktop

Question 14
Write a PySpark DataFrame to a CSV file.

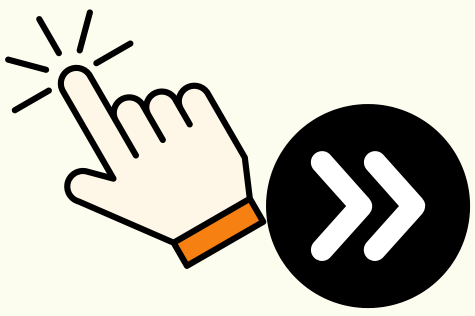
Link : https://www.linkedin.com/posts/karthik-kondpak_pyspark-practice-day-14-activity-7244582272589291520-KNcO?utm_source=share&utm_medium=member_desktop

Question 15
How would you use the rank() function in PySpark to rank employees based on their salary within their department?

Link : https://www.linkedin.com/posts/karthik-kondpak_pyspark-practice-day-15-activity-7246360630000005120-UZXj?utm_source=share&utm_medium=member_desktop

Question 16
Task: Perform a simple arithmetic operation on DataFrame columns in PySpark.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-16pdf-activity-7246723028250730496-InYX?utm_source=share&utm_medium=member_desktop



Swipe for more

https://www.seekhobigdata.com/

Question 17

PySpark coalesce Interview Question
Given a PySpark DataFrame with 8 partitions, use the coalesce function to reduce the number of partitions to 4.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-17-activity-7247102724306018305-RrK4?utm_source=share&utm_medium=member_desktop

Question 18

You are given a DataFrame containing customer transactions. The columns are customer_id, transaction_date, and amount. Write a PySpark code to calculate the following:
The total transaction amount for each customer.
The average transaction amount for each customer.
The number of transactions made by each customer.
Filter out customers who have made more than 5 transactions.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-18-activity-7247466908109701120-GQIP?utm_source=share&utm_medium=member_desktop

Question 19

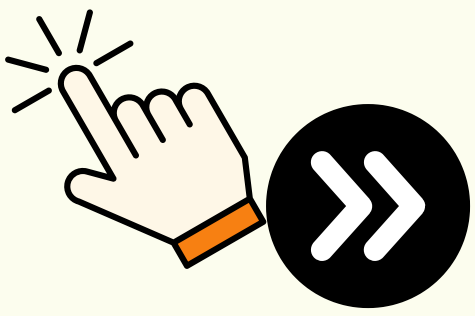
Different Ways to Read Data into PySpark
In PySpark, there are various ways to read data from different sources such as CSV, JSON, Parquet, ORC, and databases like MySQL.

Link :https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-19-activity-7247802859117813760-5w3u?utm_source=share&utm_medium=member_desktop

Question 20

How do you create a SparkSession in PySpark? What are its main uses?

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-20-activity-7248184945561477121-TY5M?utm_source=share&utm_medium=member_desktop



Swipe for more

https://www.seekhobigdata.com/

Question 21
Filter customers whose names start with 'A' and display their details.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-21-activity-7248567477029249024-oWpz?utm_source=share&utm_medium=member_desktop

Question 22
To calculate the percentage of total salary that each employee contributes to their respective department.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-22-activity-7248902885881618433-ETSM?utm_source=share&utm_medium=member_desktop

Question 23
Replace the department name "Finance" with "Financial Services" in the DataFrame:

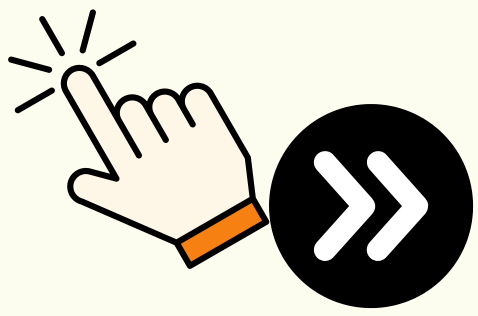
Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-23-activity-7249259743809044481-yldu?utm_source=share&utm_medium=member_desktop

Question 24
Optimize PySpark jobs for performance (tuning configurations, parallelism, etc.).

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-24-ugcPost-7249630110281019392-AZQK?utm_source=share&utm_medium=member_desktop

Question 25
Calculate the correlation between columns in a PySpark DataFrame.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-25-activity-7249983386163765248-34JD?utm_source=share&utm_medium=member_desktop



Swipe for more

https://www.seekhobigdata.com/

Question 26
Handle time-series data in PySpark

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-26-activity-7250346918432718848-QLm3?utm_source=share&utm_medium=member_desktop

Question 27
PySpark – How to Update Nested Columns?

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-27-activity-7251070550347816960-j1hZ?utm_source=share&utm_medium=member_desktop

Question 28
Explain PySpark UDF with the help of an example.

Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-28-activity-7251444789940781058-1wbU?utm_source=share&utm_medium=member_desktop

Question 29
The given file has a delimiter ~|. How will you load it as a spark DataFrame?

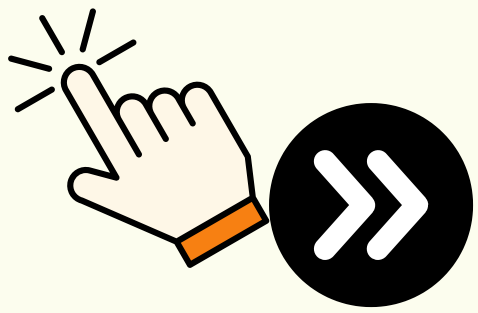
Link : https://www.linkedin.com/posts/karthik-kondpak_pysparkpracticeday-29-activity-7251803915661279235-pu0X?utm_source=share&utm_medium=member_desktop

Question 30
To cover all the concepts and commands of PySpark for Data Engineering.

Link : https://www.linkedin.com/posts/karthik-kondpak_pyspark-day-30-activity-7252316602010935296-07Ad?utm_source=share&utm_medium=member_desktop

Swipe for more

https://www.seekhobigdata.com/



If you find this helpful like and share

https://www.seekhobigdata.com/

+91 99894 54737