SODS

B.Sc III SEMESTER

PAPER: Hadoop for Big Data

PYSPARK ASSIGNMENT - I

Faculty: Ms.Manpreet Kaur Bhatia **Submission Date: 28th Oct, 2023**

Create a CSV file named "person_data.csv" with 25 entries under the following column headings: 'S.No', 'Name', 'Age', 'Occupation', and 'Education'. After creating the CSV file, initiate a Spark session and use it to read the "person_data.csv" file.

The task is to create the CSV file, start a Spark session, and load the data from the CSV file into a Spark DataFrame.

Question 1: Filtering and Selecting Data Write PySpark code to perform the following tasks:

- Create a PySpark DataFrame named **df** using the provided dataset and schema.
- Filter the DataFrame to select only the rows where age is greater than or equal to 25.
- Select only the "name" and "age" columns from the filtered DataFrame.
- Display the resulting DataFrame.

Question 2: Aggregating Data Write PySpark code to perform the following tasks:

- Calculate the average age of all the individuals in the DataFrame.
- Calculate the maximum age in the DataFrame.
- Calculate the minimum age in the DataFrame.
- Display these aggregate results.

Question 3: Grouping and Aggregating Data Write PySpark code to perform the following tasks:

- Create a new DataFrame by adding a "department" column to the existing DataFrame with values "HR" for id 1 and 2, and "IT" for the rest.
- Group the new DataFrame by the "department" column.
- Calculate the average age for each department.
- Display the resulting DataFrame with department and its corresponding average age.