PYSPARK

S.R.TAANUSRI

26.12.2023

**ETL(Extract, Transform, Load):**

* The provided PySpark code demonstrates an ETL process for handling Big Data.
* PySpark, the Python API for Apache Spark, is highlighted for its performance, ease of use, scalability, and rich ecosystem.
* The ETL workflow involves extracting data from an external CSV file, performing various transformations using PySpark functions, and saving the transformed data to an external CSV file.
* Key transformations include concatenating first and last names, calculating net salary, adding an age column, filtering data by age, calculating average salary by age, and sorting the data by age.
* The code showcases the power of PySpark in handling large-scale data processing tasks with simplicity and scalability.

**CODE:**

