**Sunkari Manohar**

**Assessment 4 – 26/12/2023**

Today we started with sample programs in pyspark.

**Parallelize ():** Here we used Parallelize() method to create an RDD for the program.

* First we created a spark session.
* Imported modules.

And after this discussed about ETL.

**ETL :**

ETL stands for

* **Extract**
* **Transform**
* **Load**

Below we performed ETL on sample **CSV** data.

**OUTPUT:**

First I loaded the data using

**Spark.read.csv(“filename.csv”)** and stored it in a variable.

Added a column called “ net salary” using

Withcolumn(“column\_name”,condition)

Here I added a full name column by concatenating the first name and last name of the users.

Here the salary is generated using rand()function and added as a new column as net\_salary.

In the same way added “age” column using withcolumn()built in function.

**Using filter:** Filter will filter the resultant data set based on the condition.

Here we are filtering data where age > 30.

**GroupBy()** with respect to net salary and age columns and below is the output

**Order by:** it Orders the data based on some condition.