**PYSPARK EXERCISES 6 (5.9.2024)**

**1. \*\*Extract\*\*:**

**#Load the employee data from a CSV file containing the following columns: `name`, `age`, `gender`, and `salary`.**

import pandas as pd

**# Create a sample CSV data**

data = {

    "name": ["John", "Jane", "Mike", "Emily", "Alex"],

    "age": [28, 32, 45, 23, 36],

    "gender": ["Male", "Female", "Male", "Female", "Male"],

    "salary": [60000, 72000, 84000, 52000, 67000]

}

df = pd.DataFrame(data)

**# Save the DataFrame as a CSV file**

csv\_file\_path = "/content/sample\_people.csv"

df.to\_csv(csv\_file\_path, index=False)

**# Confirm the CSV file is created**

print(f"CSV file created at: {csv\_file\_path}")

df\_sample\_people=spark.read.format("csv").option("header","true").option("inferSchema","true").load(csv\_file\_path)

df\_sample\_people.show()

from pyspark.sql.functions import col,avg

**2.\*Transform\*\*:**

**#\*\*Filter\*\*: Only include employees aged 30 and above in the analysis.**

df\_filtered = df.filter(col("age") >= 30)

df\_filtered.show()

**#Add New Column\*\*: Calculate a 10% bonus on the current salary for each employee and add it as a new column (`salary\_with\_bonus`).**

df\_with\_bonus = df\_sample\_people.withColumn("salary\_with\_bonus", col("salary") \* 1.10)

df\_with\_bonus.show()

**#\*Aggregation\*\*: Group the employees by gender and compute the average salary for each gender.**

df\_avg\_salary = df\_with\_bonus.groupBy("gender").agg(avg("salary").alias("avg\_salary"))

df\_avg\_salary.show()

**3.Load\*\*:**

**#Save the transformed data (including the bonus salary) in a Parquet file format for efficient storage and retrieval.**

**#Ensure the data can be easily accessed for future analysis or reporting.**

output\_path = "/path\_to\_save/transformed\_employee\_data.parquet"

df\_with\_bonus.write.mode("overwrite").parquet(output\_path)

print(f"Transformed data saved to: {output\_path}")