Lab 03 - sohamsd9

Soham Deshkhaire

Table of contents

import os import sys import pandas as pd from pyspark.sql import SparkSession from pyspark.sql.functions import col from pathlib import Path

spark = SparkSession.builder.appName(“JobPostingsAnalysis”).getOrCreate() df = spark.read.option(“header”, “true”).option(“inferSchema”, “true”).option(“multiLine”,“true”).option(“escape”, “"”).csv(“lightcast\_job\_postings.csv”)

OUTPUT\_DIR = “solution” Path(OUTPUT\_DIR).mkdir(exist\_ok=True) company\_df = df.select( col(“COMPANY”).alias(“COMPANY\_ID”),  
col(“COMPANY\_NAME”).alias(“COMPANY\_NAME”), col(“COMPANY\_RAW”).alias(“COMPANY\_RAW”), col(“COMPANY\_IS\_STAFFING”).alias(“COMPANY\_IS\_STAFFING”) ).distinct() company\_df.toPandas().to\_csv(f”{OUTPUT\_DIR}/company.csv”, index=False)

job\_postings\_df = df.select( col(“ID”).alias(“JOB\_ID”), col(“TITLE\_RAW”).alias(“TITLE\_RAW”), col(“TITLE\_CLEAN”).alias(“TITLE\_CLEAN”), col(“POSTED”).alias(“POSTED”), col(“EXPIRED”).alias(“EXPIRED”), col(“SALARY\_FROM”).alias(“SALARY\_FROM”), col(“SALARY\_TO”).alias(“SALARY\_TO”), col(“MIN\_YEARS\_EXPERIENCE”).alias(“MIN\_YEARS\_EXPERIENCE”), col(“MAX\_YEARS\_EXPERIENCE”).alias(“MAX\_YEARS\_EXPERIENCE”), col(“SKILLS”).alias(“SKILLS”), col(“SPECIALIZED\_SKILLS”).alias(“SPECIALIZED\_SKILLS”), col(“SOFTWARE\_SKILLS”).alias(“SOFTWARE\_SKILLS”), col(“EMPLOYMENT\_TYPE\_NAME”).alias(“EMPLOYMENT\_TYPE”), col(“COMPANY”).alias(“COMPANY\_ID”) ).distinct() job\_postings\_df.toPandas().to\_csv(f”{OUTPUT\_DIR}/job\_postings.csv”, index=False)

job\_location\_df = df.select( col(“ID”).alias(“JOB\_ID”),  
col(“CITY\_NAME”).alias(“CITY”), col(“STATE\_NAME”).alias(“STATE”), col(“COUNTY\_NAME”).alias(“COUNTY”), col(“LOCATION”).alias(“LOCATION”) ).distinct() job\_location\_df.toPandas().to\_csv(f”{OUTPUT\_DIR}/job\_location.csv”, index=False)

soc\_details\_df = df.select( col(“ID”).alias(“JOB\_ID”),  
col(“SOC\_2021\_2”).alias(“SOC\_2”), col(“SOC\_2021\_2\_NAME”).alias(“SOC\_2\_NAME”), col(“SOC\_2021\_3”).alias(“SOC\_3”), col(“SOC\_2021\_3\_NAME”).alias(“SOC\_3\_NAME”), col(“SOC\_2021\_4”).alias(“SOC\_4”), col(“SOC\_2021\_4\_NAME”).alias(“SOC\_4\_NAME”), col(“SOC\_2021\_5”).alias(“SOC\_5”), col(“SOC\_2021\_5\_NAME”).alias(“SOC\_5\_NAME”) ).distinct() soc\_details\_df.toPandas().to\_csv(f”{OUTPUT\_DIR}/soc\_details.csv”, index=False)

lot\_details\_df = df.select( col(“ID”).alias(“JOB\_ID”),  
col(“LOT\_CAREER\_AREA”).alias(“LOT\_CAREER\_AREA”), col(“LOT\_CAREER\_AREA\_NAME”).alias(“LOT\_CAREER\_AREA\_NAME”), col(“LOT\_OCCUPATION”).alias(“LOT\_OCCUPATION”), col(“LOT\_OCCUPATION\_NAME”).alias(“LOT\_OCCUPATION\_NAME”), col(“LOT\_SPECIALIZED\_OCCUPATION”).alias(“LOT\_SPECIALIZED\_OCCUPATION”), col(“LOT\_SPECIALIZED\_OCCUPATION\_NAME”).alias(“LOT\_SPECIALIZED\_OCCUPATION\_NAME”) ).distinct() lot\_details\_df.toPandas().to\_csv(f”{OUTPUT\_DIR}/lot\_details.csv”, index=False)

naics\_details\_df = df.select( col(“ID”).alias(“JOB\_ID”), col(“NAICS\_2022\_2”).alias(“NAICS2”), col(“NAICS\_2022\_2\_NAME”).alias(“NAICS2\_NAME”), col(“NAICS\_2022\_3”).alias(“NAICS3”), col(“NAICS\_2022\_3\_NAME”).alias(“NAICS3\_NAME”), col(“NAICS\_2022\_4”).alias(“NAICS4”), col(“NAICS\_2022\_4\_NAME”).alias(“NAICS4\_NAME”), col(“NAICS\_2022\_5”).alias(“NAICS5”), col(“NAICS\_2022\_5\_NAME”).alias(“NAICS5\_NAME”), col(“NAICS\_2022\_6”).alias(“NAICS6”), col(“NAICS\_2022\_6\_NAME”).alias(“NAICS6\_NAME”) ).distinct() naics\_details\_df.toPandas().to\_csv(f”{OUTPUT\_DIR}/naics\_details.csv”, index=False)

spark.stop()